Extracting Detailed Tobacco Exposure From The Electronic Health Record

By

Travis John Osterman

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in
Biomedical Informatics
August 11, 2017
Nashville, Tennessee

Approved

Josh Denny, M.D., M.S.
Mia Levy, M.D., Ph.D.
Pierre Massion, M. D.

To Laura, Owen and Gavin. Thank you for your patience, encouragement, and support.

ACKNOWLEDGEMENTS

I would like to thank the National Library of Medicine and the Conquer Cancer Foundation for supporting my biomedical informatics training and research (LM007450, R01-LM010685). My appreciation and thanks also go to the Department of Biomedical Informatics faculty and students for their support. Specifically, I thank Cindy Gadd and Rischelle Jenkins for creating and maintaining a rigorous and enlightening curriculum and to my master’s committee Josh Denny, Mia Levy, and Pierre Massion, for supporting my education and broader career development. In addition, I would like to thank Wei-Qi Wei, Dara Mize, Julie Wu, and Lisa Bastarche who directly contributed to this work. Finally, I would like to thank my parents, wife, and children for supporting me through this program.

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Lung cancer is the leading cause of cancer-related mortality worldwide with 1.2 million deaths annually.[1,2] Five year survival rates for patients diagnosed with lung cancer are 16.8%, largely due to most patients having metastatic disease at time of diagnosis.[3] The goal of lung cancer screening is to detect disease at an earlier stage leading to improved survival.[4] The National Lung Cancer Screening Trial (NLST) demonstrated a 20% reduction in the relative risk of lung cancer-associated mortality through annual screening of high risk patients.[5,6] Based in part on those results, the United States Preventative Services Task Force (USPSTF) recommends annual screening for patients age 55-80 who have 30 or more pack-years (PY) of smoking history, and who have not quit smoking more than 15 years.[7]

There is current a deficit in providers' ability to identify and refer patients to lung cancer screening programs. In 2015, a survey of 212 primary care physicians, 53% did not know the criteria for lung cancer screening.[8] Providers knowing three or more components were seven times more likely to screen.[8]

Clinical decision support (CDS) has been shown to improve screening rates in breast, cervical, and colorectal cancers.[9–21] Criteria for these cancer screenings depend on age, gender, and interval from last screening intervention which are data elements that may be structured within the electronic health record (EHR). By leveraging those elements, systems could be developed to prompt providers to consider screening eligible patients. Some potential screening candidates may require additional, potentially unstructured information to modify their cancer screening, such as those with hereditary cancer syndromes.[22] Fortunately, these individuals are typically a small portion of the population. Thus, cancer screening CDS can rely primarily on structured data from within the EHR as a method to reach the majority of the at risk population. Lung cancer screening introduces an additional challenge in that the data needed to determine an individual's eligibility requires knowledge of total tobacco exposure and if they have quit smoking, for how long. Neither the 30 pack-year requirement or the quit duration requirement (no more than 15 years) are typically structured in the EHR. Even when structured smoking data are present, augmenting those data through other text extract mining methods are often necessary to improve performance such as natural language processing (NLP).[23]

Natural Language Processing

The foundation of natural language procession (NLP) as a method for text extraction began in the 1950's and was initially conceived to be separate from the field of information retrieval which focused primarily on efficient indexing and searching of large documents.[24] Natural language processing focuses on the meaning and concepts of terms and phrases and extends from the early work of Chomsky who, in 1956, published a seminal paper describing three models for describing language.[25] Chomsky's work continued eventually leading to creation of Backus-Naur Form (BNF) notation which continues to be used in computer science to validate the syntax of computer programming languages.[26] As the theoretical framework of natural language processing was being described by Chomsky, Ken Thompson released a utility, *grep*, for the UNIX operating system which was the first computer program to leverage regular expressions.[27] Grep and regular expressions continue to be integral components of NLP and text processing.

In the 1960's the linguistic string parser (LSP) was created at New York University mapping grammar rules to terms.[28,29] Using LSP in the 1970's, medical terminology was incorporated into grammatical rules and sentence structure.[30] This served as the basis for processing clinical notes to determine context-aware topics and sentiments.

As the field of NLP progressed, statistical and other analytics methods such as support vector machines, hidden Markov models, conditional random fields, and N-grams were incorporated leading to the convergence of NLP and the previously separate field of information retrieval.[31–34] These divergent approaches reflect a common methodology in NLP which is subdividing the problem into smaller tasks and then applying a method tuned to each sub-task is often more successful than using the same approach throughout a complex problem. From this arises the concept of pipelined NLP frameworks such as the General Architecture for Text Engineering (GATE)[35] and later the Unstructured Information Management Architecture (UIMA).[36]

Since clinical text offers many challenges not found in other sources, multiple systems have specifically been developed to specifically extract medical concepts from clinical systems including MedLEE[37], MetaMap[38], cTAKES[39], KnowledgeMap.[40] While each of these frameworks provides validated and powerful tools to extract and map clinical concepts to standardized vocabularies such as Unified Medical Language System and/or SNOMED, simpler approaches such as keyword matching and strictly rules based systems also continue to be actively developed, typically for more defined tasks.[41–43] For the specialized task of extracting detailed smoking history including pack-years, a rules-based, regular expression approach, is appropriate.

Tobacco Extraction

Previous studies showed that smoking status could be extracted from narrative text into general classes of “ever smoker,” “never smoker,” and “former smoker.”[44] This specific task gained recognition due to the 2006 informatics for integrating biology and the bedside (i2b2)

Shared Task Smoking Status Discovery challenge.[45] This challenge posed the question if a patient's smoking status could be automatically determined. Specifically, 11 teams submitted 23 submissions to classify patients into five categories 1) unknown 2) non-smoker 3) smoker 4) current smoker 5) past smoker. Systems were scored by F-measure with the highest ranking system, submitted by Cheryl Clark and team from the MITRE Corporation with a final F-measure of 0.9.[46,47] The competing systems employed a variety of methods to determine classification, but there were several common approaches including use of machine learning (only one system used a rules-only approach), use of an ensemble of methods (most often rule-based and machine learning), and use of an initial filter to try to identify individuals whose status is unknown before proceeding forward to more complex classification tasks.

Having learned from all these systems, the team from Mayo presented an updated system three years later in 2009 built upon UIMA and leveraging cTAKES.[48] This updated system reported a micro-average F-measure of 96.7%.[48] At the same time, cTAKES was being released for the first time as an open source system.[49] This allowed the system which was developed at Mayo Clinic with local data to be tested at other medical centers. In 2012 Liu and colleagues, updated the Mayo Clinica system and validated it using the Vanderbilt University Medical Center (VUMC) SD.[44,50] Their analysis varied from the initial i2b2 challenge as they were only able to categorize patients into four categories (past smoker, current smoker, non-smoker, and unknown) due to the level of annotation available. The system was trained on 200 patients and validated on a separate 200 patients.[44] Testing occurred at the document level and both document level and patient level results were reported. Document level micro averaged F-measure for the module was 0.89 with current smoker precision scoring lowest at 0.76 and past smoker precision scoring best at 1.0. Overall, this study sets a benchmark for tiered binary classification of smoking status (never vs. ever smoker; if ever smoker, then current or past smoker) and also illustrates that the classification difficulty is not uniform across tasks.

The data extracted from NLP systems developed to identify smoking status have been leveraged to answer more complex questions using EHR data such as characterizing patients with asthma exacerbations.[51,52] Typically, these systems utilize a classification system similar to that previously described in the i2b2 challenge in which individuals are assigned either “ever,” “never,” or “unknown” smoking status and then individuals with the “ever” label are divided into “current” and “former.” Further, these tools often are able to leverage semi-structured medical records and utilize section tagging to extract social history or possibly have a semi-structured field for smoking history.[53] Each of these methods of conditioning the input helps to improved performance measures.

More recently De Silva and colleagues attempted to apply a machine learning approach utilizing rate, duration, and quantity of tobacco smoking. Smoking is quantified in pack-years where 1 pack-year equals 1 year of smoking 1 pack per day of cigarettes. One pack-year is thus equal to smoking approximately 7305 cigarettes – assuming 1 package contains 20 standard cigarettes and 365.24 days in a year.[54] De Silva’s system utilized a pipeline that begins with a set of regular expressions to identify high probably smoking terms and creates a context window of 50 characters on either side. The authors then normalize written number such as “one” and “two” to “1” and “2,” respectively including some typical fractions such as one half to “0.5.” Using these data, the authors are able to map frequency, duration, and quit time phrases as input to two support vector machines which classify based on the i2b2 labels previously described. At least in part by adding the concepts of frequency, duration, and quit time, the authors were able to achieve an F-measure of 0.95 on the i2b2 set. This analysis is also helpful as it contains a concurrent analysis of the system's perform on a data set of patient notes from the United States Department of Veteran's Affairs EHR which scored an F-measure of 0.872. This difference illustrates the challenges in comparing smoking extraction systems by performance metrics on different corpora of notes.

In the past few years, there has been increasing interest in capturing smoking and other social or environmental exposures during clinical encounters. The is most evidenced by the 2014 Institute of Medicine (IOM) report *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*.[55] In this report, the IOM recommended two tobacco-related screening questions be documents for patients each visit and be followed up if positive. It is therefore reasonable to believe that in the future factors such as smoking status may become fully structured within the EHR to be compliant with recommendations like the aforementioned. Wang and colleagues attempted to answer this question in 2016 by comparing the three different systems for identifying smoking status: patient-provided information, International Classification of Disease, Ninth Revision (ICD-9) code, and NLP.[56] They found that NLP performed best for any single system. ICD-9 codes did not meaningfully add any information to the other two methods. Finally, the combination of patient-provided information along with NLP was superior to either alone. To date, this author is unaware of a direct comparison between structured smoking data and NLP-extracted smoking data from a clinical system. For now, NLP continues to play an active role in determining a patient's smoking status for secondary research on large data sets.

Phenome Wide Association Studies (PheWAS)

Numerous studies have been performed in attempt to find associations between genetic factors, often single nucleotide polymorphisms (SNPs), and expression of disease. Genome wide association studies (GWAS) describe a method for scanning multiple, often hundreds of thousands, of genetic factors and performing statistical analysis against the expression of a single disease to quantify whether each genetic factor is a risk (or possibly protective) for that disease.[57] As the number of SNPs that can be simultaneously processed on a single chip has risen and the cost per sample has decreased, the interest and number of GWAS has grown to 10,210 SNP-phenotype associations.[58] Published studies on at least 100,000 SNPs and all SNP-disease associations with p-values less than 1.0×10^{-5} from GWAS since 2008 are stored in the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) GWAS Catalog.[59]

The combination of growth in EHR data along with the popularity of GWAS gave rise to the first phenome wide association study (PheWAS) in 2010.[60] PheWAS builds on the same fundamentals as GWAS but whereas GWAS uses many SNPs as independent variables and a disease expression as a dependent factor to compute statistical association, PheWAS utilizes EHR data to compute statistical association for numerous diseases (or phenotypes) against multiple SNPs. In the initial PheWAS, Denny and colleagues used EHR data to define 776 disease populations for 6005 individuals using ICD-9 codes. Five SNPs with previously reported associations were then studied in the method previously described. They were able to replicate 4 of 7 SNP-disease associations and identify 19 previously unknown statistical associations. This study was seminal is illustrating the secondary use of EHR data, validating the use of research grade phenotypes from EHR data, and created a new method for discovering SNP-disease associations.

The electronic MEdical Records and GEnomics (eMERGE) Network is a consortium that has developed and validated electronic phenotype algorithms for genetic studies using the EHR, often pooling analyses between sites to achieve greater statistical power.[61] The group was formed in the fall of 2007 and consisted initially of five sites and now includes 13 sites with the mission of advancing knowledge and best practices of linking biobanks with EHR data.[61]

The PheWAS methodology is not limited to disease-gene associations. In 2012, the first non-genetic PheWAS was performed by Warner and colleagues. Their analysis showed the method's ability to discover associations between laboratory values (white blood cell count in that analysis) and EHR-derived phenotypes.[62] Warner's analysis showed a significant association between *Clostridium difficile* infection and white blood cell count. More recent studies have expanded laboratory-phenotype associates of 21 laboratory measures.[63] In addition to each of these studies showing the non-genetic applications of PheWAS, they also highlight the utility in using continuous variables instead of binary SNP variables.

The accessibility of performing PheWAS analyses is also more widely available as currently there are packages available for both the R statistical language and python.[64,65] The combination of constant addition of genetic data to biobanks, the use for non-genetic association studies, and the availability of modules to aid in analysis ensure PheWAS will continue to be used to discover novel associations going forward.

Genotype by Environment Interaction Studies (G×E)

Genotype by environment interaction (G×E) studies have a long history beginning in behavioral health, but are playing an increasing role in basic science research. These studies can help to describe the complex interaction between the genetic and environmental factors contributing to disease. Genotype by environment interactions have been shown to be frequently seen in model organisms.[66] Recent studies reports G×E interactions found in human monocytes and dendritic cells.[67–69] In human disease, interactions have been studied for several specific diseases including orofacial clefts, coronary artery calcification, coronary heart disease, and obesity.[70–73] Researchers studying G×E interactions need to address a few inherent issues that have been summarized previously.[74] McGue and Carey concisely summarize the three major challenges to current G×E research in behavioral health, but these apply equally to basic science or other human disease areas (Figure 1).[75]

1. most published G×E findings are based on small samples and thus a high proportion are likely to be false-positive reports
2. imprecision in the assessment of the phenotype, environment, and the genotype can significantly attenuate the power of a G×E study
3. a G×E is not an inherent property of the organism but rather a feature of a statistical model and so its identification depends on the structure of that model.

Figure 1. Major challenges of G×E research as outlined by McGue and Carey.[75]

The exact method of determining interactions can be from a variety of methods including adding an interaction term to a typical logistic regression model, using non-linear Bayesian kernels, or hierarchical Bayesian models.[76,77] With increased access to genetic data and potentially structured exposure data, the interest in investigating G×E interactions is likely to see continued increase in interest.

CHAPTER II

SMOKING HISTORY AND PACK-YEAR EXTRACTION SYSTEM (SHAPES)

Introduction

The challenge in electronically identifying patients for lung cancer screening is that national screening guidelines require individuals are 55-80 years of age, have smoked at least 30 pack-years of cigarettes, and have not quit smoking for more than 15 years.[5] One pack-year is defined as having smoked 1 pack per day for one year (approximately 7305 cigarettes based on 20 cigarettes per package). The latter two eligibility criteria are not easily translated into computable algorithms (pack-years and quit time). While natural language processing (NLP) has been widely studied in extracting smoking status, these systems are insufficient to determine eligibility for lung cancer screening.[46,48,78] The purpose this chapter is to describe the author's initial efforts to extract and calculate quantitation smoking history in the form of pack-years in the "demonstration project." The following sections then describe expansion of this project into the Smoking History And Pack-year Extraction System (SHAPES) which removes dependencies required by the demonstration project, improves efficiency, and determines duration of time quit smoking, as applicable. These systems are developed to help support lung cancer screening clinical decision support, but quantitative tobacco exposure is useful in many biomedical informatics areas.

Demonstration Project: Introduction

The purpose of the demonstration project is to determine whether tobacco exposure could be accurately extracted and quantified from free text of clinical notes. Classification of smoking status has been extensively studied[44–46,79] and was not the primary focus of the demonstration project. The main purpose of the demonstration project was to determine feasibility of identifying individuals for lung cancer screening. Individuals with either an explicit or implicit non-zero pack-year history were of most interest. An example of implicit non-zero pack-year history would be documentation of a smoking rate (often expressed in packs per day of cigarettes) and the duration of smoking (often expressed in years). When rate is normalized to packs per day and duration to years, the product of smoking rate and duration equals smoking quantity expressed in pack-years. Explicit documentation of non-zero smoking quantity is also common in the electronic health record (EHR). An example of an explicit smoking quantity statement is "has smoked 20 pack-years."

Demonstration Project: Methods

This study used Vanderbilt University Medical Center's (VUMC) de-identified electronic health record (EHR), the Synthetic Derivative (SD).[50] A training set of 250 individuals with a history of smoking was identified from the SD using a previously validated algorithm.[44] The training set was reviewed and used to develop a rule-based pack-year extraction system built using python 2.7 and regular expressions. Building on success from previous smoking status applications, we implemented a tiered, rule-based, system using regular expression. The first rules determined whether the sentence discussed smoking. The second set of rules determined whether pack-year calculation is possible. Smoking rates and durations were converted to standard units. A third set of rules then extracted rate, duration, and/or total pack-years. Finally, a smoking quantity is calculated in the units of pack-years.

A validation set of 1000 individuals was selected randomly from the SD without enrichment for smoking status. The unstructured free text social history from each clinical note, when present, was extracted from the note using a local section tagger and stored.[80] The most common, non-blank, social history for each patient was considered the individual's true social history for purposes of the demonstration project. Note-level annotation or performance was not performed. Sixteen individuals with no available social histories were removed from the set prior to analysis.

Each individual's social history was manually reviewed by two physicians to quantify tobacco usage in pack-years which served as the ground truth. Physician assessment was compared to prediction from the rules-based system and performance measures of precision, recall, and F-measure were reported. Tobacco quantity in pack-years was analyzed using Pearson correlation and root mean square error (RMSE). An exact binomial test was used to analyze individuals meeting United States Preventative Services Task Force (USPSTF) screening criteria as previously described and reported using a p-value.

Demonstration Project: Results

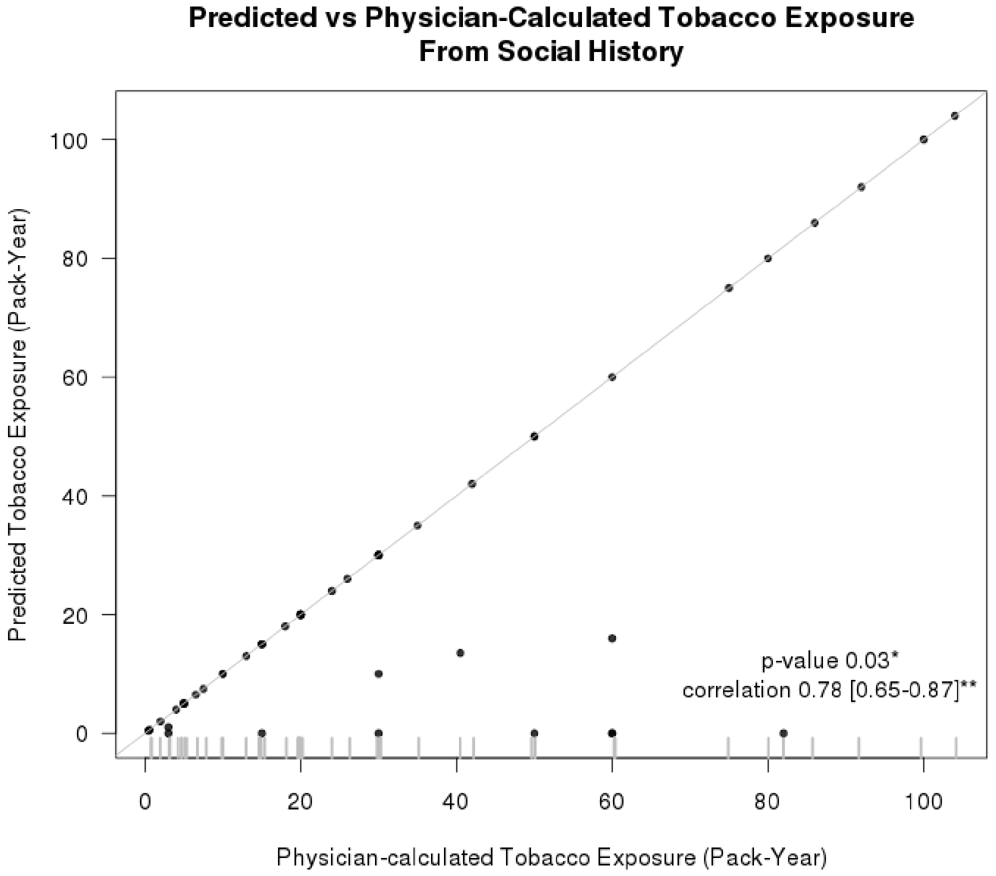


Figure 2. Predicted versus physician-calculated tobacco exposure in pack-years from social history derived from clinical text.

The demonstration projects' pack-year determination exactly matched physician chart review for 46 of 53 social histories (Pearson correlation of 0.78, 95% confidence interval [0.65 – 0.87], Figure 2)[81]. In a planned subset analysis, 14 of 19 of individuals meeting USPSTF screening criteria on physician chart review were identified (binomial exact test, one-tail α = 0.05, p-value 0.03)[81].

Demonstration Project Discussion

This small project showed that it was possible extract meaningful tobacco quantities from clinical text. There were several limitations that would potentially hinder adoption, however.

Primarily, this was a small sample size and few individuals in the validation set qualified for lung cancer screening (less than 2%) as the set was not conditioned for age. Also, the demonstration project relied on well-structured note sections so that just the social history could be processed. Some social histories may be missed by the section tagger. Tobacco use narratives outside the social history were also not included. Next, the demonstration project used a crude method of determining patient-level smoking truth in a longitudinal record (assuming the most common smoking history was true). For some problems, this may be accurate; however, this is less robust than determining tobacco exposure at the note-level and allowing future researchers to synthesize the longitudinal tobacco history with a method appropriate for a given research question. Most importantly, the demonstration project did not account for individuals that may be ineligible for screening due to having quit smoking more than 15 years ago. Nevertheless, the project met its primary goal which was to determine feasibility of a system that could extract quantitative smoking details from clinical text.

Preliminary testing confirmed that this process was highly parallelizable and capable of processing hundreds of notes each second on modest server hardware which makes real-time or large scale electronic identification of patients for lung cancer screening attainable.

Smoking History And Pack-year Extraction System (SHAPES): Introduction

Building off work done in the demonstration project, here we present the Smoking History And Pack-year Extraction System (SHAPES) to address the issues highlighted in the demonstration project. As SHAPES would be expected to extract duration of time since quitting smoking, an early design decision was to split the extraction system into the tasks of determining 7 detailed smoking values, 3 binary values (if a never-smoker, if an ever-smoker if has quit) and 4 continuous variables (rate of smoking, duration of smoking, smoking quantity, and time quit smoking), Table 1.

Binary Smoking Variables	Continuous Smoking Measures
1. Never smoker 2. Ever smoker 3. Former Smoker (has quit)	4. rate of smoking 5. duration of smoking 6. smoking quantity 7. time quit smoking

Table 1. Seven variables extracted by SHAPES.

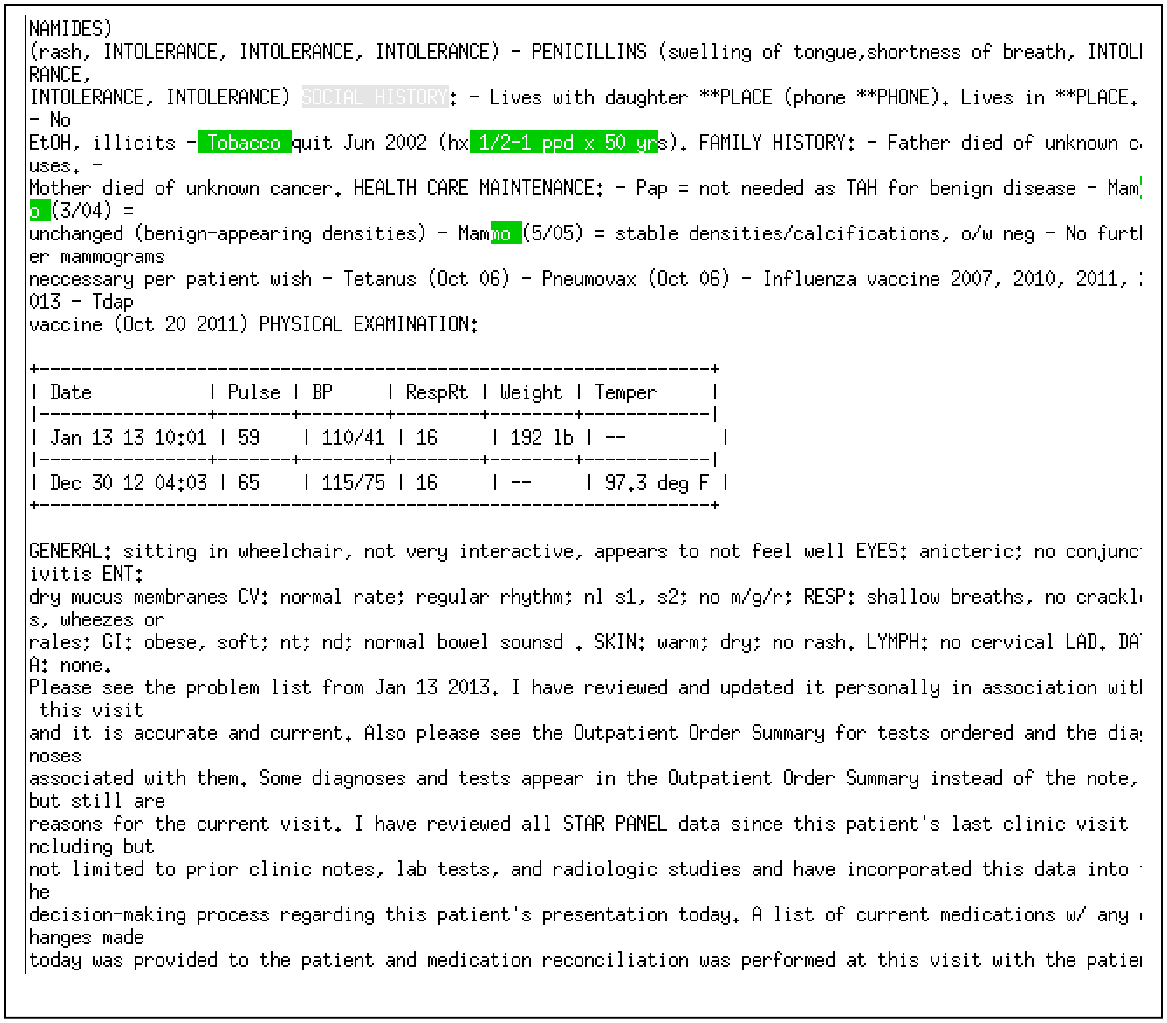
A separate task would then take the results from each of these and determine the truth for that note. SHAPES was built to determine only note-level truth. Several strategies can be employed to determine true smoking exposure at the patient level depending on the researchers' desire for precision or recall.

By processing all notes in an individual's clinical record, SHAPES needed a more robust initial filter than designed in the demonstration project. A ruled based, pre-screening process was a key component of many of the i2b2 natural language processing (NLP) smoking status extraction systems.[46] Two other problems introduced by processing every note in the patient record was 1) full notes are much longer than social histories and 2) gold standard classification would take much longer. After reviewing the available note annotation systems, we determined that to classify sufficiently large training and validation sets, a new custom augmented review and annotation system (ARAS) would need to be developed.

SHAPES: Methods

To establish a training set, 261 patient records were randomly selected from the VUMC SD[50], containing 9573 clinical notes. In attempt to get a longitudinal exposure to the medical center with a variety of notes by date, author, and department, training set selection was limited to patients within the VUMC medical home. Patients with this designation tend to have a longer relationship with VUMC including receiving primary care at VUMC and not uncommonly seeing multiple sub-specialty groups. Once selected, the following values for each note were saved to a tab-delimited file: note identifier; patient identifier; individual date of birth; note date of service; note type; and free, unstructured, text from note. These values were saved to a tab-delimited file and used for iterative rule creation within SHAPES.

SHAPES: Methods: Augmented Review and Annotation System (ARAS)



```
NAMIDES)
(rash, INTOLERANCE, INTOLERANCE, INTOLERANCE) - PENICILLINS (swelling of tongue,shortness of breath, INTOLE
RANCE,
INTOLERANCE, INTOLERANCE) SOCIAL HISTORY: - Lives with daughter **PLACE (phone **PHONE). Lives in **PLACE.
- No
EtOH, illicits - Tobacco quit Jun 2002 (hx 1/2-1 ppd x 50 yrs). FAMILY HISTORY: - Father died of unknown ca
uses. -
Mother died of unknown cancer. HEALTH CARE MAINTENANCE: - Pap = not needed as TAH for benign disease - Mamm
o (3/04) =
unchanged (benign-appearing densities) - Mammo (5/05) = stable densities/calcifications, o/w neg - No furth
er mammograms
neccessary per patient wish - Tetanus (Oct 06) - Pneumovax (Oct 06) - Influenza vaccine 2007, 2010, 2011, 2
013 - Tdap
vaccine (Oct 20 2011) PHYSICAL EXAMINATION:

+-----------------------------------------------------------------+
| Date            | Pulse | BP      | RespRt | Weight | Temper     |
|-----------------+-------+---------+--------+--------+------------|
| Jan 13 13 10:01 | 59    | 110/41  | 16     | 192 lb | --          |
|-----------------+-------+---------+--------+--------+------------|
| Dec 30 12 04:03 | 65    | 115/75  | 16     | --     | 97.3 deg F |
+-----------------------------------------------------------------+

GENERAL: sitting in wheelchair, not very interactive, appears to not feel well EYES: anicteric; no conjunct
ivitis ENT:
dry mucus membranes CV: normal rate; regular rhythm; nl s1, s2; no m/g/r; RESP: shallow breaths, no crackle
s, wheezes or
rales; GI: obese, soft; nt; nd; normal bowel sounsd . SKIN: warm; dry; no rash. LYMPH: no cervical LAD. DAT
A: none.
Please see the problem list from Jan 13 2013. I have reviewed and updated it personally in association with
 this visit
and it is accurate and current. Also please see the Outpatient Order Summary for tests ordered and the diag
noses
associated with them. Some diagnoses and tests appear in the Outpatient Order Summary instead of the note,
but still are
reasons for the current visit. I have reviewed all STAR PANEL data since this patient's last clinic visit i
ncluding but
not limited to prior clinic notes, lab tests, and radiologic studies and have incorporated this data into t
he
decision-making process regarding this patient's presentation today. A list of current medications w/ any c
hanges made
today was provided to the patient and medication reconciliation was performed at this visit with the patien
```

Figure 3. Example of the contextual highlighting interface, ARAS, used to aid rapid physician classification.

A command line interface review and annotation system was developed using python to assist expert reviewers in efficiently labelling and extracting information from patient notes.

ARAS was iteratively developed as the training set was being annotated. Each note is displayed and following values were then prompted from the user: ever smoker, current smoker, years quit, packs per day, duration in years, pack years, and comments (Figure 3). The default option for each of these fields is unknown (-1). A numeric entry is expected for all fields except comments. The fields ever smoker and current smoker expect -1, 0, or 1 for unknown, no, and yes, respectively. The fields years quit, packs per day, duration in years, and pack years, expect -1 for unknown, 0, or a positive decimal. The reviewed uses the enter key to end entry for that field and begin entry for the next field. After having the opportunity to enter comments, the reviewer is given the option to confirm her choices or to delete the entry and begin again on the same record. The reviewer can exit the system at any point by entering 'q' or 'quit' in any field and all work is saved to a binary object and to a tab-delimited file to resume later. A CLI was chosen to maximize speed of data entry. Full classification can be done using just a numerical keypad (if no comments are entered).

Additionally, the annotation time for each note was saved. Several measures were implemented in order to improve accuracy, reduce annotation time, and reduce annotation burden of redundant data.

At VUMC as records are populated into the SD, all patient identifiers are removed and dates are shifted. One artifact of this process is that entries appear in the record that are difficult to read and slow the manual review process (ex: **AGE[in 40's]). Within ARAS, instances of these artifacts are replaced by a more human readable version (ex: 40).

To alert reviewers to areas of the clinical text that may contain smoking-related text, ARAS uses a rule-based system to highlight words or phrases matching a defined set regular expression-to-color tuples (ex: ['quit(?!e)','red'] highlights quit in red text). A custom header is also printed on the screen prior to displaying the note which included note type, date of service, and the individual’s age at time service.

During the annotation process, the rule set to define a context window containing all needed smoking-related terms was iteratively refined. Between the note text and the reviewer prompts, an estimated context window was displayed for reviewer convenience. If pertinent data were found within the text, not included in the context window, the rules to extract the context window were refined. The context window extraction for the ARAS was tuned for recall at the cost of precision. The previously described highlighting was also applied to the context window text.

As clinical notes were not limited to history and physical documents, many of the notes were short free text notes, post-operative notes, etc. with no mention of any useful smoking information. In order to further reduce annotation time, any note with no identified context window was flagged an automatically labeled with appropriate values including a comment that the note was automatically annotated. The first 3000 of automatically annotated notes in the training set were then reviewed by a single reviewer for false negatives and none were found.

It is not uncommon for the same social history or smoking string to occur in multiple notes. In order to improve annotation time, each context window string was hashed. In the event that a matching hash is found, that note is populated with the values previously entered by the reviewer along with a comment regarding automatic annotation. Context windows that matched via this mechanism are thus annotated without expert intervention.

SHAPES Methods: Training

Rules for SHAPES were iteratively created by cyclical process of 1) applying SHAPES to a portion of the training set, 2) calculating performance statistics, 3) evaluating failures, and 4) modifying or creating new rules. Cycles were repeated as many times as needed, generally to target an F-measure of 0.90. When the system was performing sufficiently well, the portion of training set included was increased. To avoid over-fitting, failures were treated as a class of problems instead of a single case. For instance, the first encounter of “since 1960” required that the system be able to calculate the time between dates extracted from text and the date of the note. In this situation, efforts were made to include similar, unseen cases, such as “starting in 1960.” This allowed for the system to anticipate phrases that were not necessarily present in the training set. The methodology is also true when several words could appear in various order prior to a variable of interest. For example, the ruleset includes a line that matches any space-separated phrase containing more than one of the following words: ‘in,the,past,for,over,more,than,only,at,least,intermitantly’ followed by a digit and time designation. This allows for expansion to phrases that were not directly seen in the training data.

To speed SHAPES’ ability to analyze notes, a user-defined number of threads process each note during the evaluation phase and push results to a thread-safe queue which merges these records into the final data structure. The resulting data structure contains the clinic note information, human annotation if present, and SHAPES annotation. Patient-level data was synthesized by combining all the note-level information. Both the patient-level and note-level data structures were then written to tab-delimited results files.

The tab-delimited results files were read by the SHAPES analysis module and processed. The human annotations were compared to SHAPES annotations and performance metrics were computed. The resulting performance metrics were saved as a tab-delimited file. Performance measures were reviewed by the author and changes were made in one of three main areas 1) correcting inaccurate human annotations 2) modifying rulesets 3) modifying SHAPES and how SHAPES applies the rulesets.

SHAPES: Methods: Assumptions

The “truth” for a given record is sometimes ambiguous so the following guidelines were created to guide human classification, rule development, and system implementation. These guidelines thus govern SHAPES’ behavior especially when dealing with ambiguous or confusing clinical text (Figure 4).

1. No outside knowledge can be applied that is not in the note text, note date, and individual birth date. Example 1: if individual is age 3 and no smoking history is present, all values are unknown even if children of that age cannot smoke. Example 2: if 5 prior notes mention the duration of smoking but that is not present in the current note, only the available information can be annotated. Example 3: if the note mentions “smoked since 1972,” and the note date is 2002, then duration is estimated at 30 years.

2. For binary (Yes/No) categories (ever smoker, never smoker, and can quantify), Yes = 1, No = 0, Unsure/unable to determine = -1

3. For continuous variables (packs per day, duration, years quit, and pack years), Unsure/unable to determine = -1; otherwise any non-negative decimal number is allowed

4. 20 cigarettes are assumed to be in each package of cigarettes

5. Smoking a pipe or marijuana preclude the individual from being a never smoker

6. Pipe, marijuana, or e-cigarette rate, duration, and quantity are not reported

7. When given a range, take higher value (ex: 1-2 ppd = 2 ppd, 10-20 years = 20 years)

8. If the individual is a documented smoker and there is no documentation regarding that individual having quit, then that individual is considered a current smoker (quit years = 0)

9. If no smoking information is present in the note: ever smoker = -1, current smoker = -1, duration = -1, rate = -1, pack years = -1, years quit = -1

10. If the individual is documented to be a never smoker, the following values are to be used: rate = 0, duration = 0, pack-years = 0, quit years = -1

11. If conflicting information exists within a single note as to whether the individual has ever smoked, error on the side of labelling the individual as an ever-smoker (ex: “no tobacco history […] smoked 1ppd x 30 years, quit 25 years ago” = Ever Smoker = 1

12. If conflicting information is present within a note on rate, duration, or quantity of tobacco, the higher value is used (example: “smoked 2 ppd x 20 years, now smoking 1/2 ppd” = 2 ppd)

13. If age is listed as “a teen,” that will be considered 13 years of age

14. If 0 packs per day is the only smoking reference, the individual is considered a never smoker

Figure 4: SHAPES assumptions and disambiguation rules.

SHAPES: Methods: Pipeline

Note processing is done in the file classifier and proceeds linearly through the steps outlined in Figure 5. Each of these steps will be reviewed below.

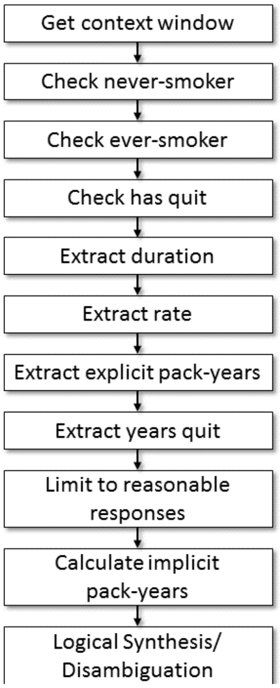

Get context window
Check never-smoker
Check ever-smoker
Check has quit
Extract duration
Extract rate
Extract explicit pack-years
Extract years quit
Limit to reasonable responses
Calculate implicit pack-years
Logical Synthesis/ Disambiguation

Figure 5. SHAPES note processing pipeline.

Determining the preferred context window is done in multiple steps. First, the note text is split by line. Each line is processes by each regular expression in the array RULES.WINDOW.INCLUSION (full ruleset included in Appendix A). If a line does not match, the next line is processed. If a line matches, the line is then processed against RULES.WINDOW.EXCLUSION and excluded if any of those rules match. If the line matches the inclusion rules but not the exclusion rules, that line along with the preceding line and proceeding line are joined. This process allows for the context window to preserve possibly important text above and below the identified smoking text. If multiple context windows exist within a note, those are all joined to create the “uncleaned context window.”

The uncleaned context window is then processed. The first step is removing quotes, extra spaces, new line characters, and html-style tags. The second step is replacing decade references with an absolute time reference. For example, a note with date 1/1/2010 and context window that contained “in the 90s” would be replaced with “x 20 years.” Next, all text references to numeric values are replaced (example “1/2” with “0.5” and “twenty” with “20”). Age and date ranges are then replaced to reflect assumption 7 (Figure 4) above (example “from 1950-1990” with “x 40 years” and “from around 19 until age 89” with “x 70 years”). Next relative dates are replaced to account for the date of service. For example, given a note written 1/1/2010, the text “since 1970” would be replaced with “x 40 years.” The next preprocessing rule applies uniform spacing around digits so that “2ppd” is converted to “2 ppd.” Relative age statements are then converted to absolute duration statements similar to those done with dates. For example, “since age 16” would be changed to “x 90 years” for a note dated 2006 and a patient who was born in 1900. Following this, non-age ranges are replaced to be consistent with assumption 7 (figure 4) above (example “1 - 2 ppd” with “2 ppd”). Finally, any remaining brackets are stripped from the text.

The resulting text is then reprocessed RULES.WINDOW.INCLUSION and a context window is created by including a user-specified number of characters to the left and right of the matches with variables RULES.WINDOW.LEFTCHARS and RULES.WINDOW.RIGHTCHARS. Care is taken to not split the text mid-word. Any overlapping context windows are appropriately merged. The context window then goes through a user-defined “cleaning” process. The purpose of cleaning is to allow for a method prior to processing that removes artifacts that might be specific to a single EHR or data warehouse. In the VUMC SD, artifacts from the de-identification process are removed via a set of rules included in this cleaning process. After cleaning, the remaining context window serves as the source for annotation by SHAPES. The uncleaned, unmodified context window is also saved throughout the processing pipeline to aid in failure analysis of the preprocessing pipeline. If present, the extracted context window is saved.

After a context window has been identified, SHAPES proceeds to process individual note-level functions that act almost entirely independently to extract the seven smoking details previously outlined (Table 1).

The first process determines whether the note being processed documents never smoking. The function relies on user-configurable variables RULES.NEVER_SMOKER.INCLUSION and RULES.NEVER_SMOKER.EXCLUSION and returns -1, 0, or 1 for unknown, have smoked at some point, and confirmed never smoker, respectively. The next sequential function checks ever smoking status and relies on user-configurable variables RULES.EVER_SMOKER.INCLUSION and RULES.EVER_SMOKER.EXCLUSION and returns -1, 0, or 1 for unknown, confirmed never smoker, and smoked at some time, respectively. Evaluating whether the note documents an individual having quit smoking is the next function which relies on the variable RULES.QUIT.HAS_QUIT and returns -1,0, or 1 similar to the previous functions.

Duration is the first continuous measure to be extracted. This function relies on variables RULES.DURATION.INCLUSION and RULES.DURATION.EXCLUSION similar to the aforementioned, but each rule is expected to have one and only one regular expression group match [ex: “(\d+)”]. For convenience a local variable ‘_d’ is provided in the rules.py file (Appendix A) that is defined as “[~]?(\d+(?:\.\d+)?)” and supports matching of most common numerical expressions that are present in a context window after preprocessing. Durations are normalized to years and returned if found, otherwise -1.

Rate is similar to duration but utilizes RULES.RATE.INCLUSION and RULES.RATE.EXCLUSION. In addition to rules processing, rates are normalized to packs per day (ex: 10 cigarettes per day = 0.5 packs per day). The rate function also attempts to determine if the individual smokes periodically and attempts to update the rate accordingly (ex: "smokes 1 pack every other day" is the same as smokes "0.5 packs per day"). If rate is not able to be determined, the function returns -1, otherwise, rate is returned in packs per day. After returning the extracted rate, a small specialized function assesses whether the rate is likely an incorrectly written pack-year expression (ex: "150ppd smoking history" is more likely "150 pack-year smoking history").

Explicit pack-year expressions are then extracted (ex: "30 pack-year history") using rules RULES.PACKYEAR.INCLUSION and RULES.PACKYEAR.EXCLUSION. Return values are congruent with rate and duration. A function here also evaluates for miswritten phases that are more likely rates (ex: "smokes 1py per day") and corrects them.

The next detailed smoking information to be extracted is duration of time quit smoking. This is processed similarly to duration of smoking including conversion of weeks and months to years. The rules applied to extract quit time are define in RULES.QUIT.INCLUSION and RULES.QUIT.YEARS_AGO. Return values are similar to rate and duration.

All rules in rules.py (Appendix A) can utilize variables for smoking ('SMK'), cigarettes ('CIG'), tobacco ('TOB'), pack ('PK'), packs per day ('PPD'), and time ('TIME'). The purpose of these is merely convenience and allows rules that are more human readable. The 'TOB' variable, for example, will match "tob" or "tobacco" but not "October" or "lactobacillus."

At the conclusion of all the aforementioned functions, a sanity check is run to ensure that the numbers extracted are sensible. The minimum and maximum allowable values are user-configurable for each category. If a value for a given smoking characteristic falls outside the defined range, the value is set back to -1 (unknown) and a message is logged as a warning.

After ensuring sane values for rate and duration, implicit pack-year quantity is determined by multiplying the rate and duration. If both implicit and explicit pack-year expressions are located within a document, the largest is saved as true (per assumption 12, Figure 4).

The final step in determining the detailed smoking information that will be returned is synthesizing the data from all the functions. This final ensemble step uses 11 conditional statements in attempts to reconcile sometimes ambiguous or conflicting data within the note. This process may end up overwriting nearly all the extracted metrics. For example, if a note documented individual as a never smoker but rate, duration, and pack-year were unknown prior to running the final function, the result would be rate, duration, and pack-year being assigned as 0 (assumption 10, Figure 4).

While the shapes module (shapes.py) does not provide a mechanism to combine note-level detailed smoking data into patient level data, this feature is provided separately both in the training module (train.py) and in the utility module (patient-level.py). The provided implementation allows for greatest values to be taken from all categories, re-calculation of pack-years as rate and duration information may be available across separate notes, and a process for ensuring the patient-level data do not conflict similar to the final step in the note assessment above.

For the purposes of presenting patient-level data here, each category of data was considered for each patient and the highest value was taken as true (ex: if one note stated the individual was a smoker, they were labelled as a smoker even if two notes claimed non-smoker). Following this process, the product of rate and duration for each patient was taken and if greater than the greatest pack-year value, the new implicit pack-year was deemed as true. This would be the situation for a record in which smoking rate and duration occur in separate notes but not together. If rate, duration, pack-years, or quit time were determined at the patient-level, the patient was considered to be have smoked. If ever-smoking and never-smoking status differed, ever-smoking status was taken.

SHAPES: Methods: Validation

In a similar methodology to the training set, 352 patient records containing 4040 notes were randomly selected from the VUMC SD for the validation set. One notable difference in selection was that individuals from the VUMC medical home were not favored. The validation set was reviewed and annotated by two physicians using the ARAS and stored in a tab-delimited file. The two physician annotation sets were compared across each of the six captured fields. Adjudication was performed on any note that did not have complete agreement across all six fields. Adjudication was performed by a third physician. The combination of notes that were in full agreement between the two primary reviewers and the final determination made by the adjudicator comprise the final validation set.

SHAPES: Results

The SHAPES training set consisted of 261 patients whose records' contained 9573 notes. All notes were for each patient were included in the set. The mean number of notes per patient was 36.7 with a median of 22 notes. The frequency of the note authors documenting never smoking status, ever smoking status, rate of smoking, duration of smoking, quantity of smoking (either implicit or explicit pack-years), and duration of quit if applicable is shown in Table 2.

	Training Set N = 9573		Validation Set N = 4040	
Category	N	Prevalence	N	Prevalence
Never smoker documented	974	10%	556	14%
Ever smoker documented	1123	12%	446	11%
Smoking rate documented	511	5%	246	6%
Smoking duration documented	430	4%	186	5%
Smoking quantity documented	488	5%	175	4%
Smoking quit time documented	735	8%	370	9%

Table 2: Frequency of note-level smoking data in training and validation sets.

Patient-level data were determined for the training and validation sets using the method outlined above and are summarized in Table 3. The smoking frequency was 38% and was the most common smoking variable at the patient-level. Only 20% of patients has pack-year data available (53% of smokers).

	Training Set N = 261		Validation Set N = 352	
Category	N	Prevalence	N	Prevalence
Never smoker documented	94	36%	132	37%
Ever smoker documented	99	38%	99	28%
Smoking rate documented	54	21%	66	19%
Smoking duration documented	48	18%	44	12%
Smoking quantity documented	52	20%	41	12%
Smoking quit time documented	94	36%	85	24%

Table 3: Frequency of patient-level smoking data in training and validation sets.

Performance statistics (sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-measure, root mean squared error (RMS), and prevalence) for each smoking variable for SHAPES when applied to the training set are presented in Table 4 along with true and false positive and negative rates.

	Never-smokers (pt)	Ever-smokers (pt)	Rate (pt)	Duration (pt)	Quantity (pt)	Years Quit (pt)
N	9573	9573	9573	9573	9573	9573
Sensitivity	0.98	1.00	0.97	0.94	0.95	0.55
Specificity	1.00	0.99	1.00	1.00	1.00	1.00
PPV	1.00	0.90	0.97	0.98	1.00	0.99
NPV	1.00	1.00	1.00	1.00	1.00	0.96
F-measure	0.99	0.95	0.97	0.96	0.97	0.71
RMS	0.05	0.21	0.08	1.21	1.94	0.98
Accuracy	1.00	0.99	1.00	1.00	1.00	0.97
TP	958	1118	497	404	463	406
TN	8595	8331	9046	9135	9085	8832
FP	4	119	16	8	0	6
FN	16	5	14	26	25	329
Prevalence	0.10	0.12	0.05	0.04	0.05	0.08

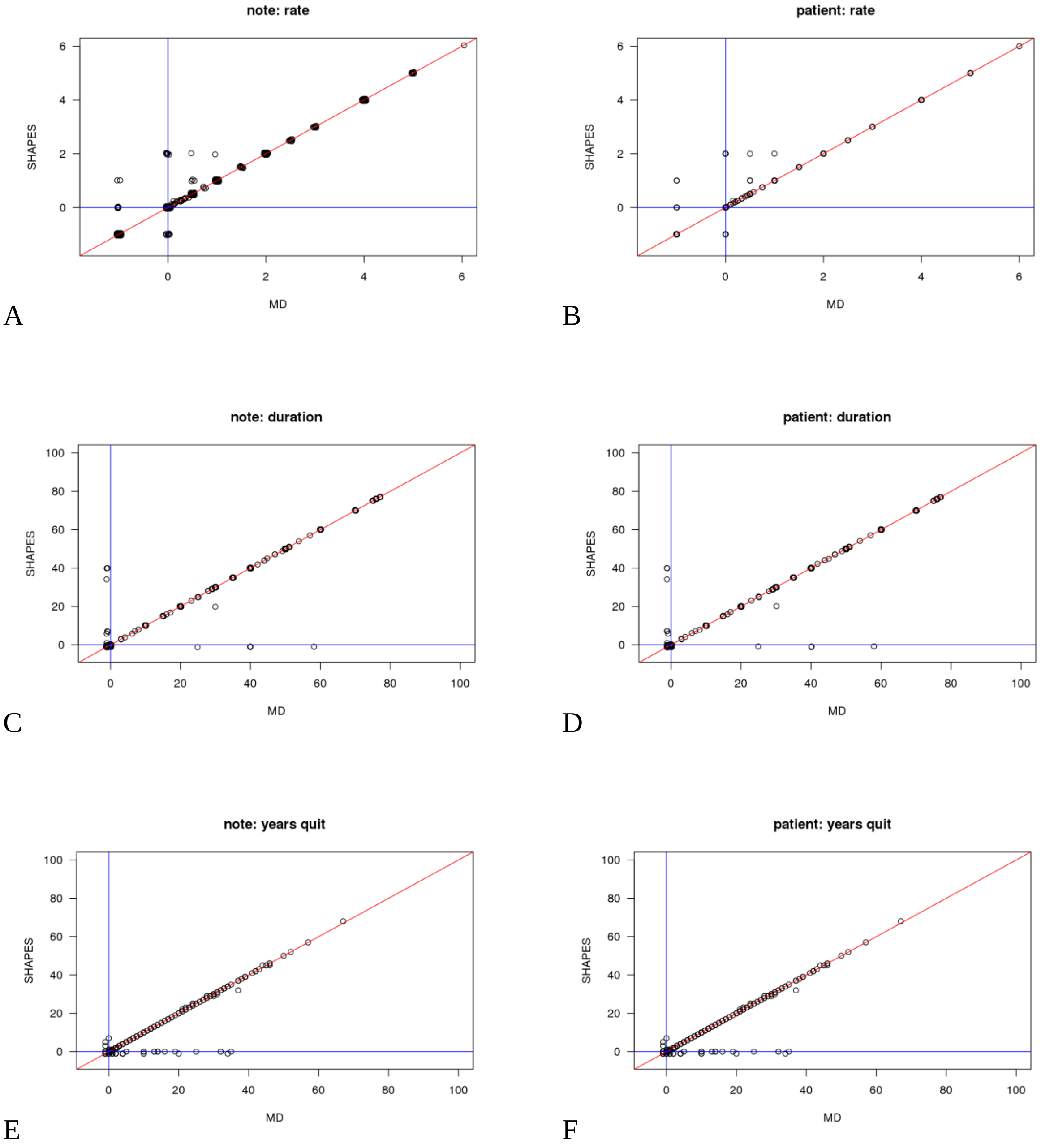
Table 4. Training set performance, note level.

The same measures are presented in Table 5 for the data at the patient-level.

	Never-smokers (pt)	Ever-smokers (pt)	Rate (pt)	Duration (pt)	Quantity (pt)	Years Quit (pt)
N	261	261	261	261	261	261
Sensitivity	0.79	0.98	0.96	0.92	0.92	0.57
Specificity	0.99	0.80	0.97	0.99	1.00	0.98
PPV	0.97	0.75	0.90	0.94	1.00	0.95
NPV	0.89	0.98	0.99	0.98	0.98	0.80
F-measure	0.87	0.85	0.93	0.93	0.96	0.71
RMS	0.29	0.51	0.24	4.02	10.83	0.97
Accuracy	0.92	0.87	0.97	0.97	0.98	0.84
TP	74	97	52	44	48	54
TN	165	130	201	210	209	164
FP	2	32	6	3	0	3
FN	20	2	2	4	4	40
Prevalence	0.36	0.38	0.21	0.18	0.20	0.36

Table 5. Training set performance, patient level.

Figure 6, below, shows a set of calibration of the training data. The left column shows note-level data with physician-review on the x-axis and SHAPES prediction on the y-axis. Points falling above the line are false-positives or over-predictions while points below the line are false-negatives or under-predictions. Points below or to the left of the blue lines denote that the smoking variable was not documented or unable to be determined. Points falling on the red line indicate agreement between the physician reviewer and SHAPES prediction. Each row in Figure 6 corresponds to the four continuous smoking variables predicted by SHAPES (rate, duration, time quit, and smoking quantity in pack-years). Axes are consistent between note-level and patient-level plots.


note: rate
SHAPES
MD
6
4
2
0
0
2
4
6
A
patient: rate
SHAPES
MD
6
4
2
0
0
2
4
6
B
note: duration
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
C
patient: duration
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
D
note: years quit
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
E
patient: years quit
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
F

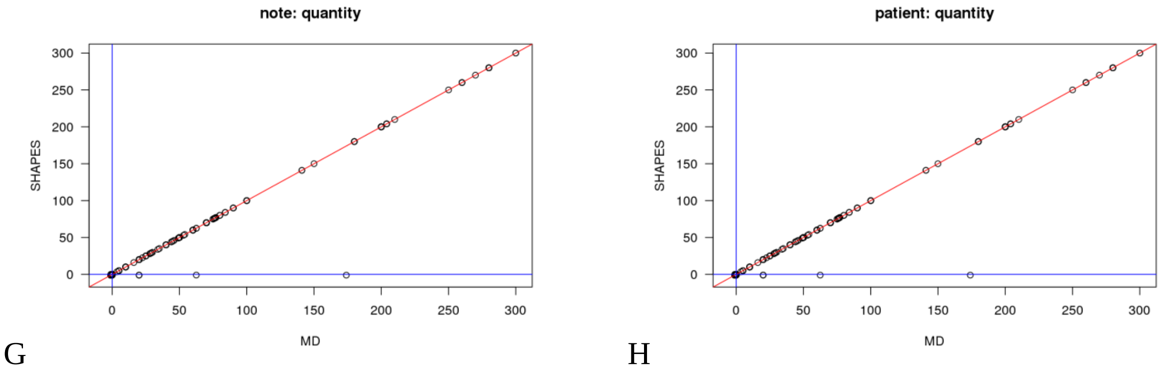


Figure 6. Training set calibration plots

The SHAPES validation set consisted of 352 patients whose records' contained 4040 notes. All notes were for each patient were included in the set. The average number of notes per patient was 11.5 with a median of 5 notes. General characteristics are summarized in Table 2, above.

Table 6 shows a summary of the note-level inter-rate reliability for the validation set. Simple agreement is shown for all variables. Cohen Kappa is included for binary variables and intraclass correlation (ICC) is showed for continuous variables.

	Reviewer 1 vs 2 including unknowns n=4040			Reviewer 1 vs 2 excluding unknowns n=variable		
	Agreement	Kappa	ICC	Agreement	Kappa	ICC
Ever	0.84	0.49		0.99	0.93	
Current	0.90	0.71		0.97	0.93	
Never	0.84	0.49		0.99	0.93	
Rate	0.86		0.53	0.95		0.30
Duration	0.86		0.84	0.93		0.98
Quantity	0.87		0.91	0.96		1.00
Quit Time	0.96		0.70	0.93		0.98

Table 6. Inter-rater reliability between two independent reviewers.

A tolerance of 0 was set for agreement between binary classifications. A tolerance of 0.1 was set as an agreement threshold for continuous variables (for example, in the case of a record stating "smokes 1 pack per week," a rate of 0.14 packs per day and 0.1 packs per day would be considered to be in agreement. Any disagreement, on any of the variables, triggered a review of the entire note for all variables. A total of 735 notes were reviewed by a third reviewer. Those values were then taken as truth for the respectively notes.

For the 735 notes that were adjudicated, inter-rater reliability was assessed between the adjudicator and each of reviewers 1 and 2 (Table 7).

	Reviewer 1 vs. Adjudicator (including unknown) n=735			Reviewer 2 vs. Adjudicator (including unknown) n=735		
	Agreement	Kappa	ICC	Agreement	Kappa	ICC
Ever	0.21	0.12		0.89	0.75	
Current	0.54	0.21		0.87	0.50	
Never	0.21	0.12		0.89	0.75	
Rate	0.27		0.07	0.93		0.86
Duration	0.27		0.44	0.91		0.81
Quantity	0.30		0.35	0.93		0.81
Quit Time	0.86		0.35	0.88		0.77

Table 7. Review vs adjudicator agreement.

The kappa scores for reviewer 1 compared to the adjudicator ranged from 0.12-0.21. The kappa scores for reviewer 2 compared to the adjudicator ranged from 0.50-0.75. Intraclass correlation values were also higher for reviewer 2 vs. adjudicator (range 0.77 – 0.86) compared to reviewer 1 vs. adjudicator (range 0.07 – 0.44).

The same analysis is shown in Table 8 except, as in Table 7 above, *n/a* values are removed prior to calculation.

	Reviewer 1 vs. Adjudicator (including unknown) n=variable			Reviewer 2 vs. Adjudicator (including unknown) n=variable		
	Agreement	Kappa	ICC	Agreement	Kappa	ICC
Ever	0.99	0.98		0.98	0.93	
Current	0.98	0.86		0.95	0.70	
Never	0.99	0.98		0.98	0.93	
Rate	0.96		0.17	1.00		1.00
Duration	0.90		0.94	0.98		0.96
Quantity	0.95		1.00	0.99		1.00
Quit Time	0.89		0.99	0.77		0.77

Table 8. Reviewer vs adjudicator agreement, ignoring NAs.

Performance statistics (sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], F-measure, root mean squared error [RMS], and prevalence) for each smoking variable for SHAPES when applied to the training set are presented in Table 9 along with true and false positive and negative rates.

	Never-smokers (pt)	Ever-smokers (pt)	Rate (pt)	Duration (pt)	Quantity (pt)	Years Quit (pt)
N	4040	4040	4040	4040	4040	4040
Sensitivity	0.89	0.95	0.59	0.32	0.34	0.29
Specificity	0.99	0.95	0.99	0.99	1.00	0.99
PPV	0.97	0.72	0.81	0.66	0.82	0.80
NPV	0.98	0.99	0.97	0.97	0.97	0.93
F-measure	0.93	0.82	0.68	0.43	0.48	0.43
RMS	0.14	0.36	0.30	3.49	5.59	2.90
Accuracy	0.98	0.95	0.97	0.96	0.97	0.93
TP	497	423	146	59	59	108
TN	3466	3426	3760	3824	3852	3643
FP	18	168	34	30	13	27
FN	59	23	100	127	116	262
Prevalence	0.14	0.11	0.06	0.05	0.04	0.09

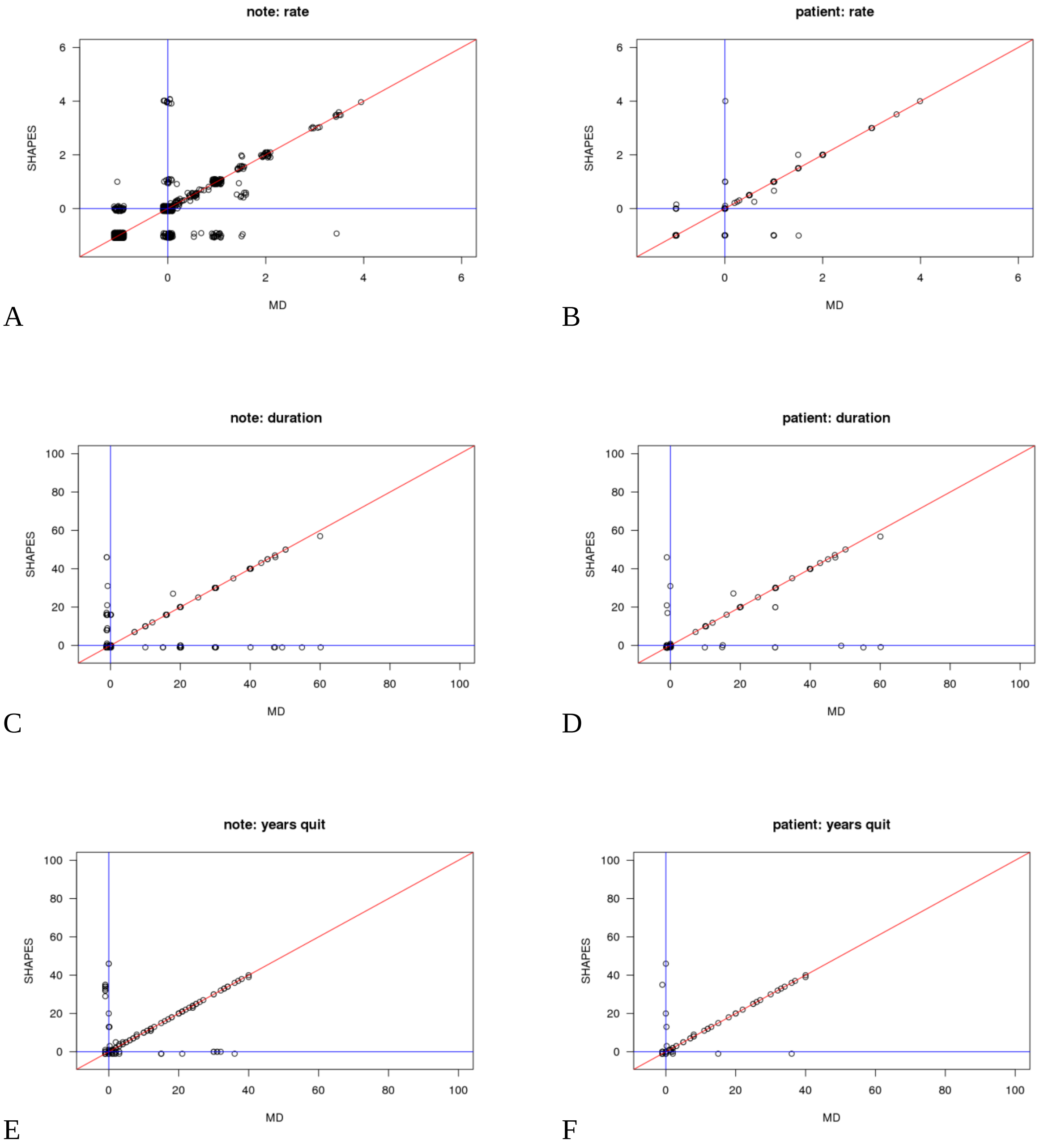
Table 9. SHAPES validation set performance (note-level data).

The same measures are presented in Table 10 for the data at the patient-level.

	Never-smokers (pt)	Ever-smokers (pt)	Rate (pt)	Duration (pt)	Quantity (pt)	Years Quit (pt)
N	352	352	352	352	352	352
Sensitivity	0.79	0.98	0.73	0.55	0.59	0.47
Specificity	0.98	0.84	0.97	0.97	0.96	0.98
PPV	0.95	0.70	0.86	0.71	0.69	0.87
NPV	0.88	0.99	0.94	0.94	0.95	0.85
F-measure	0.86	0.82	0.79	0.62	0.64	0.61
RMS	0.31	0.47	0.40	6.76	9.99	4.01
Accuracy	0.91	0.88	0.93	0.91	0.92	0.86
TP	104	97	48	24	24	40
TN	215	212	278	298	300	261
FP	5	41	8	10	11	6
FN	28	2	18	20	17	45
Prevalence	0.37	0.28	0.19	0.12	0.12	0.24

Table 10. SHAPES validation set performance (patient-level data).

Figure 7, below, shows a set of calibration of the validation data. The left column shows note-level data with physician-review on the x-axis and SHAPES predication on the y-axis. Points falling above the line are false-positives or over-predictions while points below the line are false-negatives or under-predictions. Points below or to the left of the blue lines denote that the smoking variable was not documented or unable to be determined. Points falling on the red line indicate agreement between the physician reviewer and SHAPES prediction. Each row in Figure 7 corresponds to the four continuous smoking variables predicted by SHAPES (rate, duration, time quit, and smoking quantity in pack-years). Axes are consistent between note-level and patient-level plots.


note: rate
SHAPES
MD
6
4
2
0
0
2
4
6
A
patient: rate
SHAPES
MD
6
4
2
0
0
2
4
6
B
note: duration
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
C
patient: duration
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
D
note: years quit
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
E
patient: years quit
SHAPES
MD
100
80
60
40
20
0
0
20
40
60
80
100
F

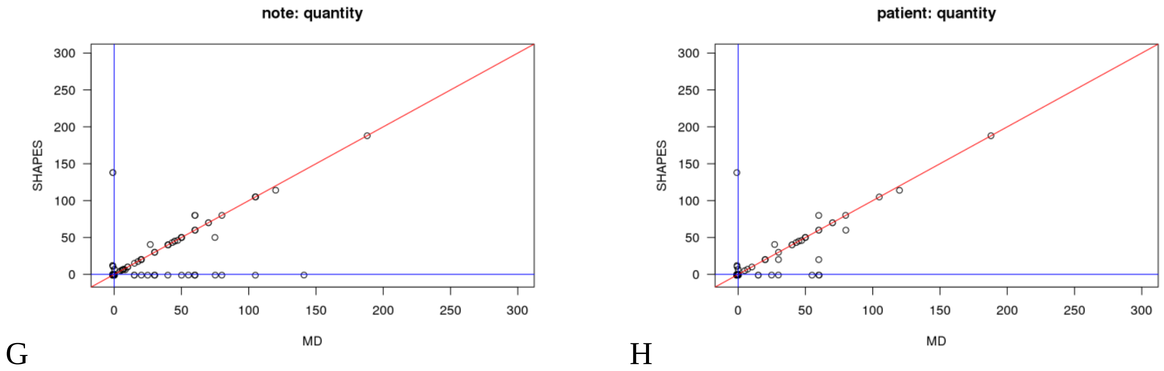


Figure 7. SHAPES validation set calibration plots for note-level (left) and patient-level (right) data

	True Eligible	True Ineligible	Total
SHAPES Eligible	16	1	17
SHAPES Ineligible	6	329	335
Total	22	330	352

Table 11. Contingency table of SHAPES lung cancer screening eligibility predictions.

When NLST criteria were applied to the to the validation set, 22 patients were eligible for low dose CT screening for lung cancer. SHAPES identified 17 patients with 1 false positive and 6 false negatives. Based on this validation set, for identifying patients qualifying for lung cancer screening, SHAPES' precision was 0.94 [0.90,0.96], recall was 0.73 [0.70, 0.75], and F-measure 0.82 (Table 10).

	True Eligible	True Ineligible	Total
SHAPES Eligible	35	12	47
SHAPES Ineligible	0	305	305
Total	35	317	352

Table 12. Contingency table of SHAPES abdominal aortic aneurysm screening eligibility predictions.

When the USPSTF abdominal aortic aneurysm (AAA) screening criteria were applied to the to the validation set, 35 patients were eligible for AAA radiographic screening. SHAPES identified all 47 patients with 12 false positives and 0 false negatives. Based on this validation set, for identifying patients for AAA screening, SHAPES' precision was 0.74 [0.73, 0.76], recall was 1.00 [0.95, 1.00], and F-measure was 0.85 (Table 11).

SHAPES: Reference Implementations

SHAPES is provided as open source software under the APACHE version 2.0 license[82] and is available for download via github.com. After download, dependencies should be installed via `pip install -r requirements`. A complete list of dependencies is provided in Appendix B.

We provide several ways to interact with SHAPES. Our expectation is that the system will be integrated into an NLP pipeline or similar system with capabilities of providing note text, note date, and patient date of birth and then capable of writing the smoking results to a file or database for use. We have been successful in leveraging python's multiprocessing module to analyze many patient records in parallel at a rate of approximately 4-5 notes/thread/second on commodity hardware. So long as the dataset is sufficiently large, the rate of scaling appears to be linear on server-grade hardware.

For users, interested in integrating the system into an existing pipeline or applying SHAPES to all patients in a large database, Figure 8 is provided as an example of importing the SHAPES module and calling it on clinical text. This is intended to be the prototypical "hello world" implementation.

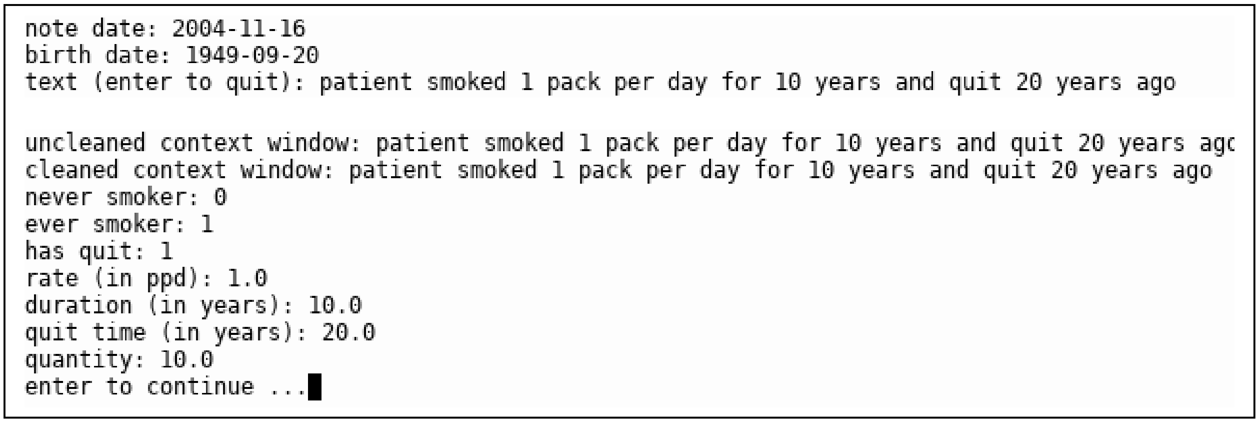
```
#!/usr/bin/python2.7 -tt
# -*- coding: utf-8 -*-
import os
import sys
temp_path =
      os.path.abspath(os.path.join(os.path.dirname(__file__),
      os.path.pardir))
sys.path.append(temp_path)
from shapes import *
import utilities as u
import shapes
sys.path.remove(temp_path)
import logging

def main():
  logging.basicConfig(filename='log/example.log',level=logging.DEBUG)
  logging.basicConfig(format='%(asctime)s %(message)s',
    datefmt='%m/%d/%Y %I:%M:%S %p')
  shapes = Shapes()
  birth_date =    '1949-09-20'
  note_date =     '2004-11-16'
  hx = """
  patient smoked 1 pack per day for ten years and quit 20 years ago
  """
  u.print_dict(shapes.get_structured_tobacco_exposure(
    hx,
    note_date,
    birth_date,
    'example')
  )

if __name__ == "__main__":
  main()
```

Figure 8. SHAPES example module implementation.

For users that wish to test the system in a smaller scale, users can execute the included `shapes_cli.py`. This 51-line program presents the user with a prompt to enter text with smoking information. After hitting enter, SHAPES processes the text and presents the user with the extracted smoking information (Figure 9). The intention of this program is to allow for rapid testing or tuning of new rules to handle specific text cases. Both uncleaned and cleaned versions of the context window are presented to aid in any needed debugging. Further debugging information is available via a log file that is written to each cycle.

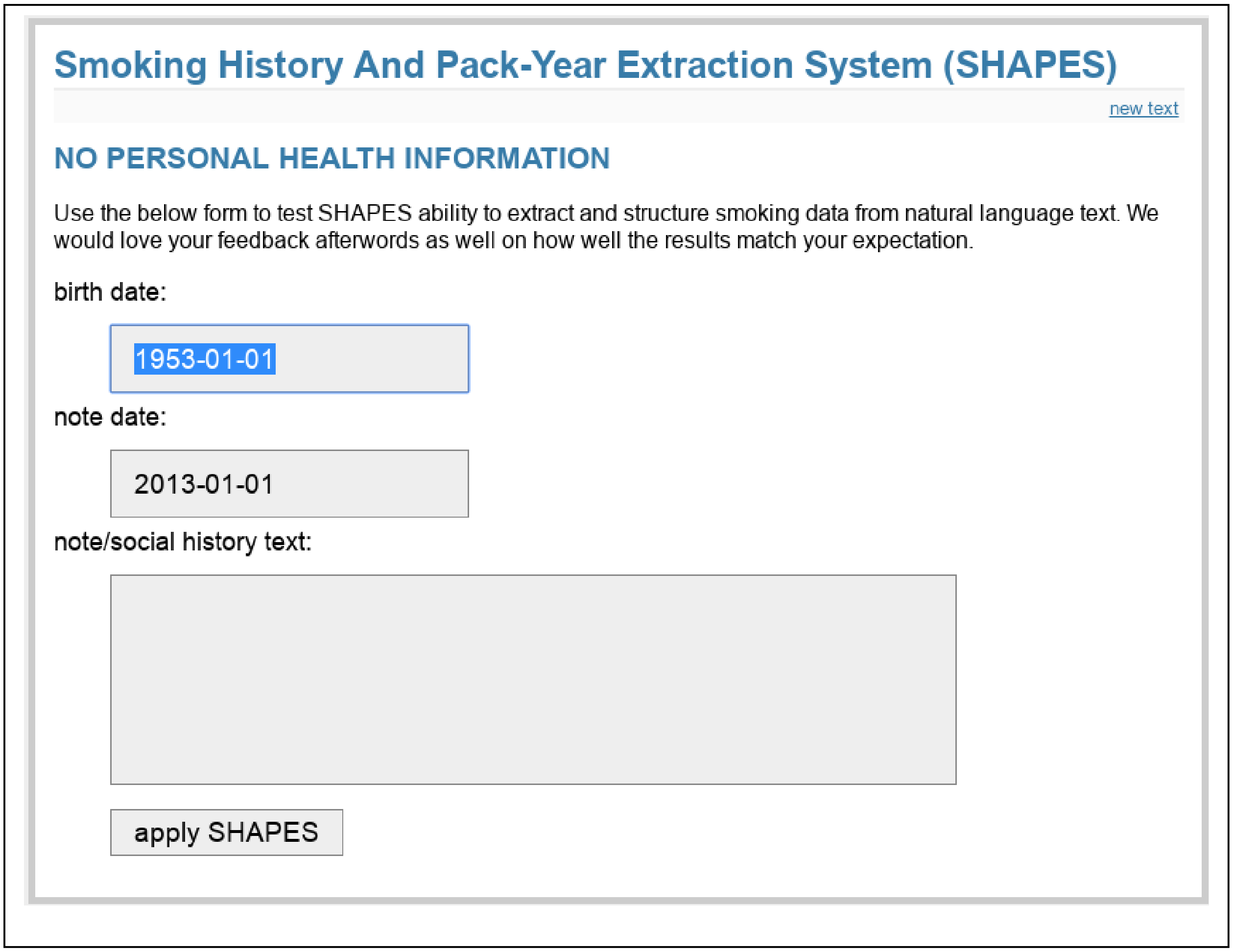


```
note date: 2004-11-16
birth date: 1949-09-20
text (enter to quit): patient smoked 1 pack per day for 10 years and quit 20 years ago

uncleaned context window: patient smoked 1 pack per day for 10 years and quit 20 years agc
cleaned context window: patient smoked 1 pack per day for 10 years and quit 20 years ago
never smoker: 0
ever smoker: 1
has quit: 1
rate (in ppd): 1.0
duration (in years): 10.0
quit time (in years): 20.0
quantity: 10.0
enter to continue ...
```

Figure 9. SHAPES command line interface

For users that may not be as comfortable with a CLI and to more easily support multiple users providing feedback simultaneously, a python web application is provided as a reference implementation to SHAPES. This program uses Flask to create a web server and host the application from the local computer (`python shapes/implementation/webapp/shapes_app.py`). The site can then be accessed via `http://localhost:8080`, Figure 10. The layout is similar to the CLI; however, the logging system is made available via the web interface and provides data on each query made, date and time of query, IP address of query, web client identifier (via cookie created at first page view), SHAPES results, and the results of the user's review of the results if applicable. While this system is provided able to run via a local web host, adapting the application to be suitable for a larger environment such as a dedicated python web server (Tornado or other) or via Apache module (mod_wsgi).



Smoking History And Pack-Year Extraction System (SHAPES)

NO PERSONAL HEALTH INFORMATION

Use the below form to test SHAPES ability to extract and structure smoking data from natural language text. We would love your feedback afterwords as well on how well the results match your expectation.

birth date: 1953-01-01

note date: 2013-01-01

note/social history text:

Figure 10. SHAPES web interface.

SHAPES: Discussion

Here we describe a natural language processing system to extract quantitative smoking history (rate, duration, quantity in pack-years, and time quit). These data may be used to support clinical decision support applications requiring more detailed smoking information such as lung cancer screening or AAA screening or may be leveraged in other research involving the EHR as smoking tobacco is a risk factor for many conditions. SHAPES was validated at the Vanderbilt University Medical Center and should be validated in other environments prior to operational use. We feel the design of SHAPES enables users to modify the rulesets easily with little fore-knowledge of the system as all rules are located in the `rules.py` file (Appendix A).

Several design decisions separate SHAPES from previous NLP-based smoking status applications. First SHAPES is built using python 2.7 which is installed by default on most major linux distributions and available for all major operating systems. When possible, python's freely available modules were leveraged (Appendix B), but SHAPES does not depend on any large NLP framework such as cTAKES or UIMA. Specifically, SHAPES was built to not require section-tagging as smoking information may be present in any portion of a clinical note. Next, SHAPES focuses on note-level truth. Deciding how one wants to handle "truth" at the patient level is a research decision. For example, some projects may require a very high precision so requiring multiple clinical notes to collaborate tobacco history may be desired which may sacrifice power. For other research questions, recall may be favored as all identified patients may undergo a full manual chart review. A project may require a small number of never-smokers only and inclusion of anyone with smoking history is costly. Each of these scenarios can be accommodated by combining note-level data.

The training and validation sets were taken from two slightly different populations (training data primarily from patients labelled as "medical home" and validation data randomly from all patients). This is reflected in the expected fewer notes per patient in the validation set. Note-level data were consistent across smoking variables. Patient-level variables did have differences in ever smoker documentation (38% of training patients, and 28% of validation patients), pack-year documentation (20% for training patients and 12% for validation patients), and quit time documentation (36% for training patients and 24% for validation patients), Tables 2 and 3. This may be due the synthesis of note-level data to patient-level data, due to fewer notes per patient, or due to higher smoking rates in the medical home population.

F-measures in the training set were greater than 0.95 for all measures except years quit. Determining the number of years quit was the most challenging variable to extract and required the largest portion of rule development time. There appear to be a few reasons for this. First, there are many ways that a provider references time quit (absolute date, month-year, year, only month, number of year, number of months, age of quit, etc.). Each one of these forms requires specific handling. A particularly challenging case to handle was "smoker from age 20 to 40," as the identifying 40 as an age quit without mention of "smoked," "stopped," or "quit" required a special case to identify smoking age ranges. A second challenge of identifying years quit is the varying distance the quit string may be from the rate or duration strings. Smoking rate and duration are often very near each other in text, the quit string may be separated by sufficient distance to fall outside the context window. A third difficulty is the case where a patient may have quit multiple times, but the expectation is to report only the most recent quit duration. A fourth difficulty in determining quit time is disambiguating quit time from other factors such as alcohol. This may be the easiest of the above to address in future work as it should be amenable to concept mapping.

In the training set, most variables had lower F-measures for note-level compared to the patient level data (never smoker 0.99 to 0.87, ever smoker 0.95 to 0.85, rate 0.97 to 0.93, duration 0.96 to 0.93, and quantity 0.97 to 0.96). F-measures for quit time in the training set were the same for note-level and patient-level data.. There are two identified factors here. First, the patient level data is a smaller sample size so maintaining F-measures of near 0.95 in a sample size of 261 is challenging. Second, if one note had a false positive with an over estimate, it would likely be carried through to the patient-level based on how note-level data is translated to patient-level data. For example, if one note erroneously identified a 300 pack-year smoking history, whereas 9 notes identified a 100 pack-year smoking history, the system was designed to favor the higher value.

Future analyses could be done to determine to optimal method of determining patient-level data based on the underlying precision and recall of the smoking variables and clinical question being addressed.

Reporting F-measures, even with a tolerance, on continuous data is a poor descriptor of correlation. Since *n/a* is also the most common value for any of the smoking variables, Pearson's correlation coefficient is also not ideally suited. The calibration plots in Figure 7 allow for a more nuanced interpretation. While the majority of predictions were correct, several rates were over-estimated, a few durations were erroneously attributed (i.e. the duration was for alcohol and not smoking), and several quit times were missed completely.

Interpreting and coding detailed smoking history from clinical notes is a challenging task even for clinical experts. To help reduce ambiguity, reviewers were provided a set of guidelines (Figure 4) to help establish agreement on ambiguous cases. Examples of ambiguity include a note that mentions the patient as being a "non-smoker" but goes on to state that the patient "quit smoking 10 years ago." In this case, the individual is deemed a former-smoker of unknown rate, duration, or quantity. In the case that the only mention is "non-smoker"; however, SHAPES determines this individual to be a lifetime never smoker with rate, duration, and quantity all equal to 0. Similarly, if an individual is felt to be younger than a generally accepted age that smoking is reasonable, age two for instance, the reviewer is instructed not to consider this. If a researcher wishes to eliminate individuals deemed too young to smoke tobacco, those rules can be applied after SHAPES.

The inter-reviewer reliability, Table 5, comparing calculating Cohen's kappa (binary categories) and intraclass correlation (ICC, continuous categories) shows the difference when *n/a*s are included or removed. For instance, Kappa for determining ever-smoker or never-smoker changed from 0.49 to 0.93 when *n/a*s were excluded. This is due to differences in reviewers' sensitivity identifying a patient as an ever-smoker. When reviewers determined the individual was an ever or never-smoker, they infrequently disagreed on which of those categories were true. This may reflect reviewer fatigue as it indicates that records were coded as *n/a* when smoking data were present. A similar pattern is seen in all categories with the exception of smoking rate. It is unclear why the ICC is worse when *n/a*s are removed. This implies that reviewers agreed that a rate was present but interpreted that rate differently.

To help determine whether the variations between reviewers was due to uniform fatigue, ambiguity in the data, or other factors, Kappa and ICC were calculated between reviewer 1 and the adjudicator and compared to those of reviewer 2 and the adjudicator, Tables 7 and 8. The former table shows that reviewer 1 appears to have missed more smoking data in the clinical notes, thus erroneously encoding categories as *n/a* as reviewer 2 had consistently higher Kappa and ICC values with the adjudicator. When *n/a*s were removed, Table 8, Kappa and ICC values were more in agreement. The difficulty in agreeing on smoking rate seems to be a difference with reviewer 1's interpretation of rate as reviewer 2 and the adjudicator has 100% agreement and ICC. Thus, we show that by establishing a set of ambiguity guidelines, Figure 4, that we are able to have reasonably agreement between physician reviewers. The differences highlighted between reviewers 1 and 2 emphasize the importance of multiple reviewers to establish a gold standard for detailed smoking history.

Note-level performance was worse in the validation set than in the training set for recall and precision across all categories. F-measures for identifying ever and never-smoking notes was 0.93 and 0.82, respectively and similar to those in the 2006 i2b2 challenge but less than those reported by the Mayo Clinic algorithm.[44,78] When compared to the training set, recall was decreased more than precision for all quantitative categories. This most likely reflects the diversity of ways smoking history is described in individual notes. Unlike the training set data, we saw an increase in performance when individual notes are combined in the way described. Table 10 shows that recall improves across all categories and precision is maintained or improved for all categories with the exception of smoking quantity. Combining notes thus allows patient-level F-measures to be higher than those for the corresponding note-level F-measures ranging from 0.61 (years quit) to 0.86 (never smoker). This difference reflects that the training set and validation set may contain a greater difference in how smoking history is expressed than anticipated. Figure 7, panel H shows that few individuals have smoking exposure over-predicted and when compared to panel G, that combining notes improves recall.

The utility of SHAPES depends on the research question. For some applications, the recall may not be acceptable. As SHAPES was developed to help support lung cancer screening, Table 11 shows a contingency table for the individuals that would be eligible for lung cancer screening based on NLST criteria. In this scenario, patient-level recall is 0.73 which is higher than either the recall for smoking quantity (0.59) or years quit (0.47) which illustrates that SHAPES does not perform equally across all ranges of smoking exposures. Current 3.9% of eligible individuals are screened for lung cancer.[83] These rates are sufficiently low such that identifying 73% of the eligible population may be worthwhile since the precision of that prediction is 94%. As screening rates improve, however, SHAPES' recall would need to improve in order to be useful.

For abdominal aortic aneurysm screening, SHAPES' recall and precision of 100% and 73%, respectively, is a good balance for a screening process if reviewed by a clinical prior to screening enrollment. It is important to note that no determination of any screening utility is made or implied by SHAPES.

SHAPES: Conclusion

Here we have shown that quantitative smoking data is difficult to extract from the EHR. The usefulness of SHAPES depends on the measures of interest and application. Attention should be paid to the difficulties in determining note-level truth even with physician reviewer when transporting SHAPES to other institutions. The author recommends users perform a local validation before widespread deployment of the system. As recall was generally lower than precision or specificity, validation on a set of ever-smokers from one of the previously mentions algorithm may be an efficient method.

CHAPTER III

SMOKING PHEWAS

Introduction

Unlike the previously described smoking classification systems, the Smoking History And Pack-Year Extraction System (SHAPES), chapter 1, was not designed to extract binary smoking status (ever vs never having smoked) so direct comparison to existing systems is challenging.[46,48,78] Since smoking classification systems are often used as a part of a larger analyses, one method for comparing SHAPES to previous smoking classification systems is within the context of a larger analysis that incorporates smoking data. As previously noted, phenome wide association studies (PheWAS) are not limited to gene-disease association analysis or binary predictors.[62,63] Thus, PheWAS provides an appropriate setting to compare the real world utility of SHAPES to a previously validated smoking status extraction system. Here we perform two PheWAS on the same population. The first study utilizes binary smoking classification. The second study uses pack-year exposure as extracted by SHAPES.

Methods

A convenience set of individuals was selected from the Vanderbilt University Medical Center (VUMC) Synthetic Derivative (SD) consisting of patients who largely participate in the VUMC medical home program[84] and thus have longitudinal records, multiple encounters with the healthcare system, a larger set of diagnosis codes, and more clinical documents. Ever-smoking vs never-smoking status was determined using a previously validated smoking classification system.[44] Pack-years were predicted for the same patients using SHAPES. International

Classification of Disease version 9 (ICD9) codes were extracted from the SD for each individual and mapped to PheCodes using the PheWAS R package.[64] Age of last contact and gender for each individual were extracted from the SD. Two separate PheWAS were run using the PheWAS R package each using age and gender as covariates with one using binary ever-never smoking status and the other using pack-year exposure, Figure 11.

$$phenotype = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 EverNeverStatus$$

$$phenotype = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 PackYearQuantity$$

Figure 11. PheWAS logistic regression using ever-never smoking status (top) vs pack-year (bottom)

A minimum of 2 ICD9 codes were required to establish a phenotype for an individual. If fewer than 20 individuals were available as cases or controls for a phenotype, that phenotype association was not tested. Results were considered statistically significant if the p-value for the association was less than 0.05 after Bonferroni correction for multiple testing (p=0.05/1816 = 2.7 x 10^{-5}).

For significant phenotypes in the SHAPES dataset, a set of cumulative probability curves was generated for a theoretical 50 year old man with a smoking history ranging from 0-200 pack-years based on the model outlined in Figure 11.

To compare the ability of the two different systems to determine significant results, a simulation was performed. The PheWAS using pack-years was repeated incrementally removing 20 individuals from the analysis. For each iteration a random group of individuals was removed and the resulting PheCode association p-values were collected. A Kolmogorov–Smirnov (KS) test was performed to calculate a D-statistic for each iteration against the binary ever-never p-values. This was repeated for a total of 1574 iterations.

Results

A total of 35,788 individuals were included in the analysis with a 15,664 unique ICD9 codes (8,628,995 in total). A total of 1816 phenotypes were mapped as described above. The ages of the population ranged from less than 1 year of age to greater than 89 (mean: 55.1, median 61).

	Binary Smoking Status	SHAPES
Individuals Processed	35,788	35,788
Smoker (> 0 PY for SHAPES)	7,515	12,591
Never Smoker (0 PY for SHAPES)	15,897	15,045

Table 13. Smoking PheWAS overview of ever-never smoking classification vs SHAPES.

One thousand, eight hundred, and sixteen association tests were performed in each group. The binary, ever/never smoking PheWAS yielded 153 associations (Appendix C) with p-values less than Bonferroni threshold, Figure 12. The top 10 association by p-value are included along with odds ratio in Table 14.

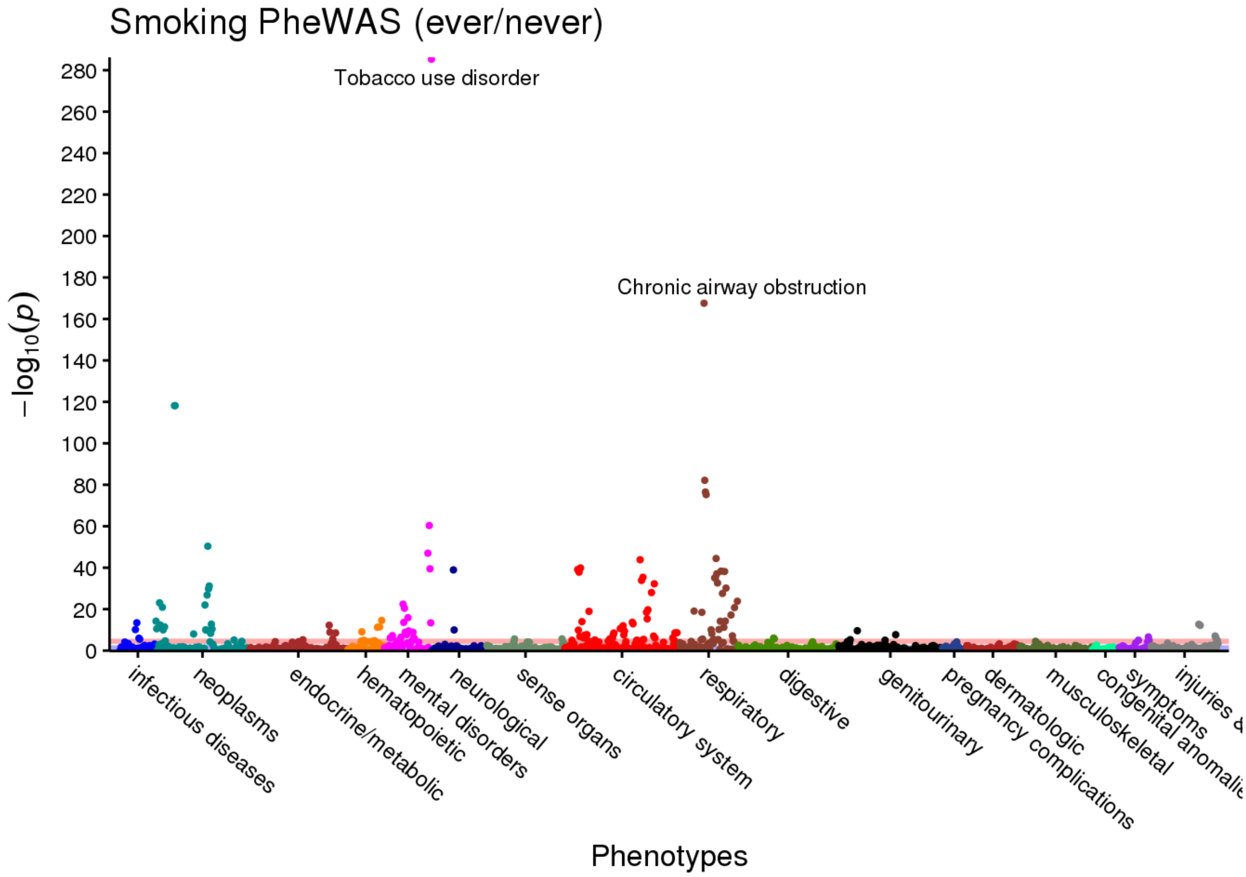


Figure 12: Smoking associations found with binary ever/never smoking status.

Description	Phecode	Code Group	p-value	OR
Tobacco use disorder	318	mental disorders	5.2E-286	6.8
Chronic airway obstruction	496	respiratory	2.6E-168	3.6
Cancer of bronchus; lung	165.1	neoplasms	6.6E-119	4.5
Cancer within the respiratory system	165	neoplasms	7.2E-119	4.4
Emphysema	496.1	respiratory	7.19E-83	6.7
Chronic bronchitis	496.2	respiratory	3.12E-77	4.6
Obstructive chronic bronchitis	496.21	respiratory	5.76E-76	5.3
Alcohol-related disorders	317	mental disorders	4.61E-61	5.0
Secondary malignant neoplasm	198	neoplasms	4.38E-51	1.8
Substance addiction and disorders	316	mental disorders	1.01E-47	3.1

Table 14. Top 10 smoking-phenotype associations by p-value for ever/never smoking status.

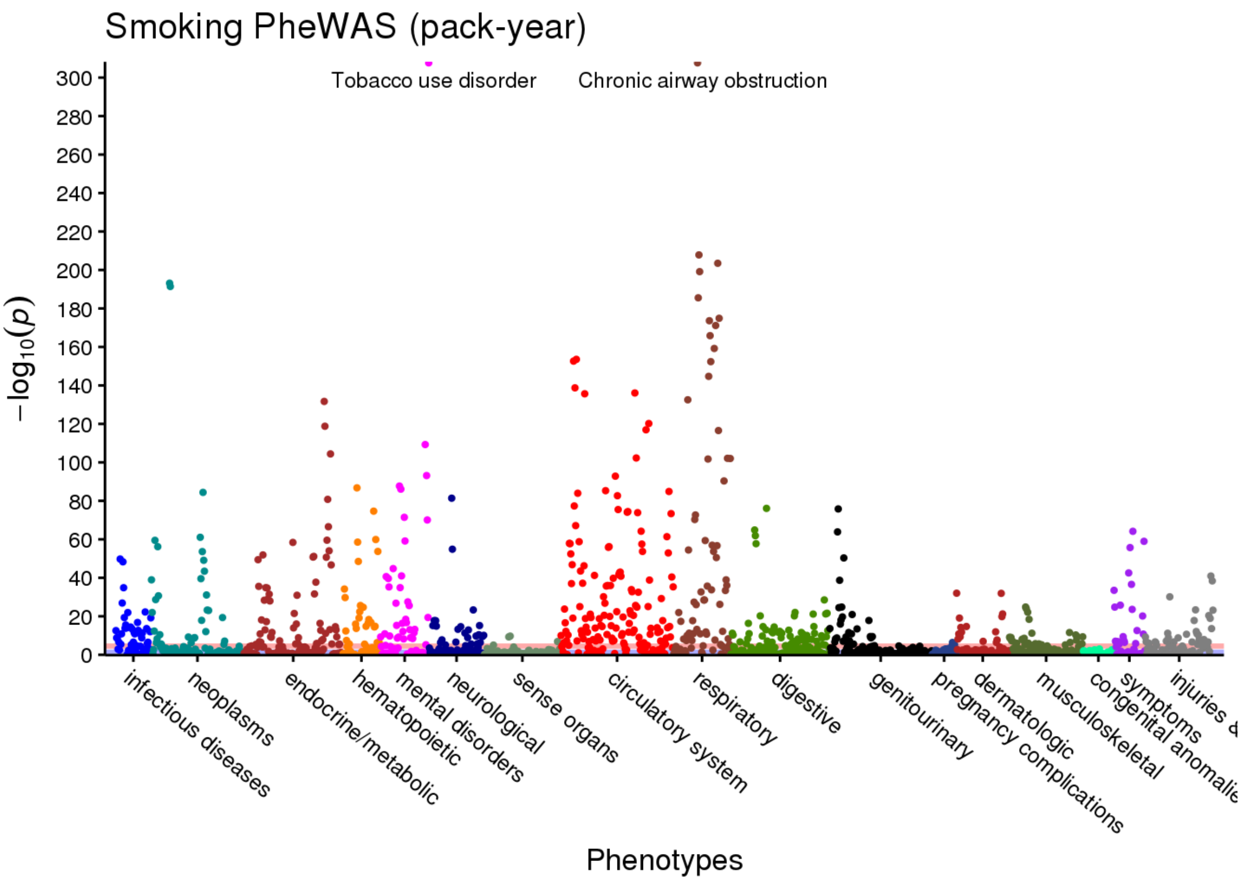
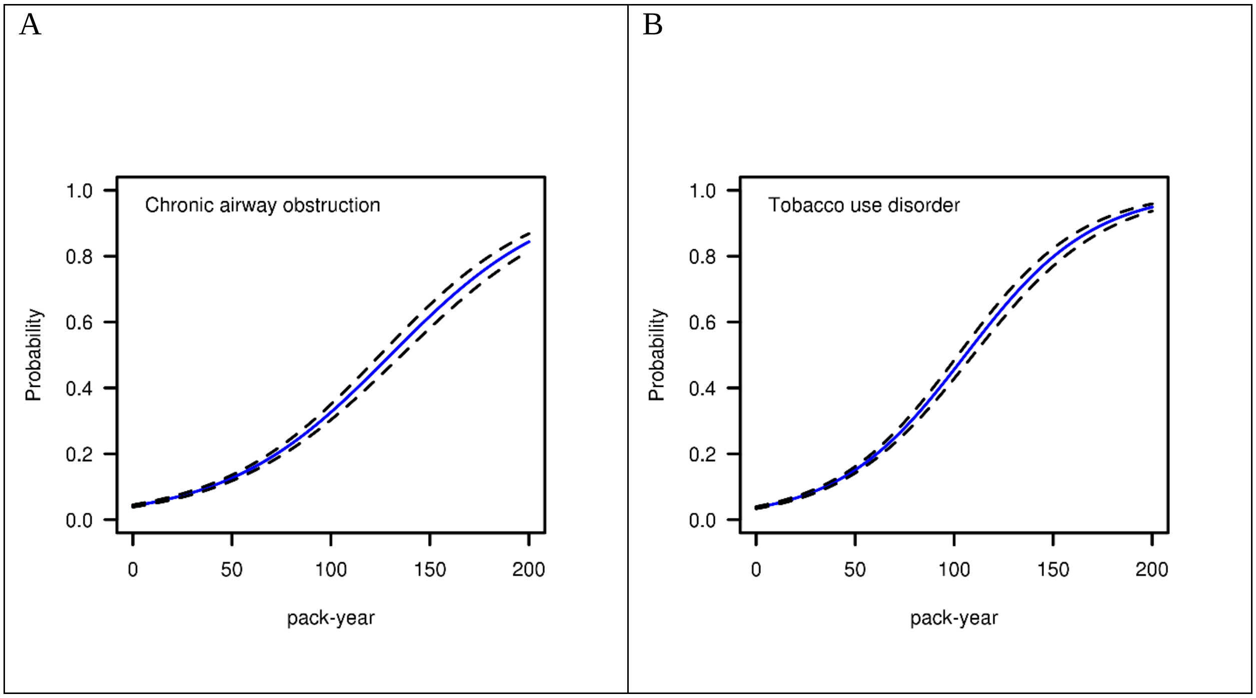


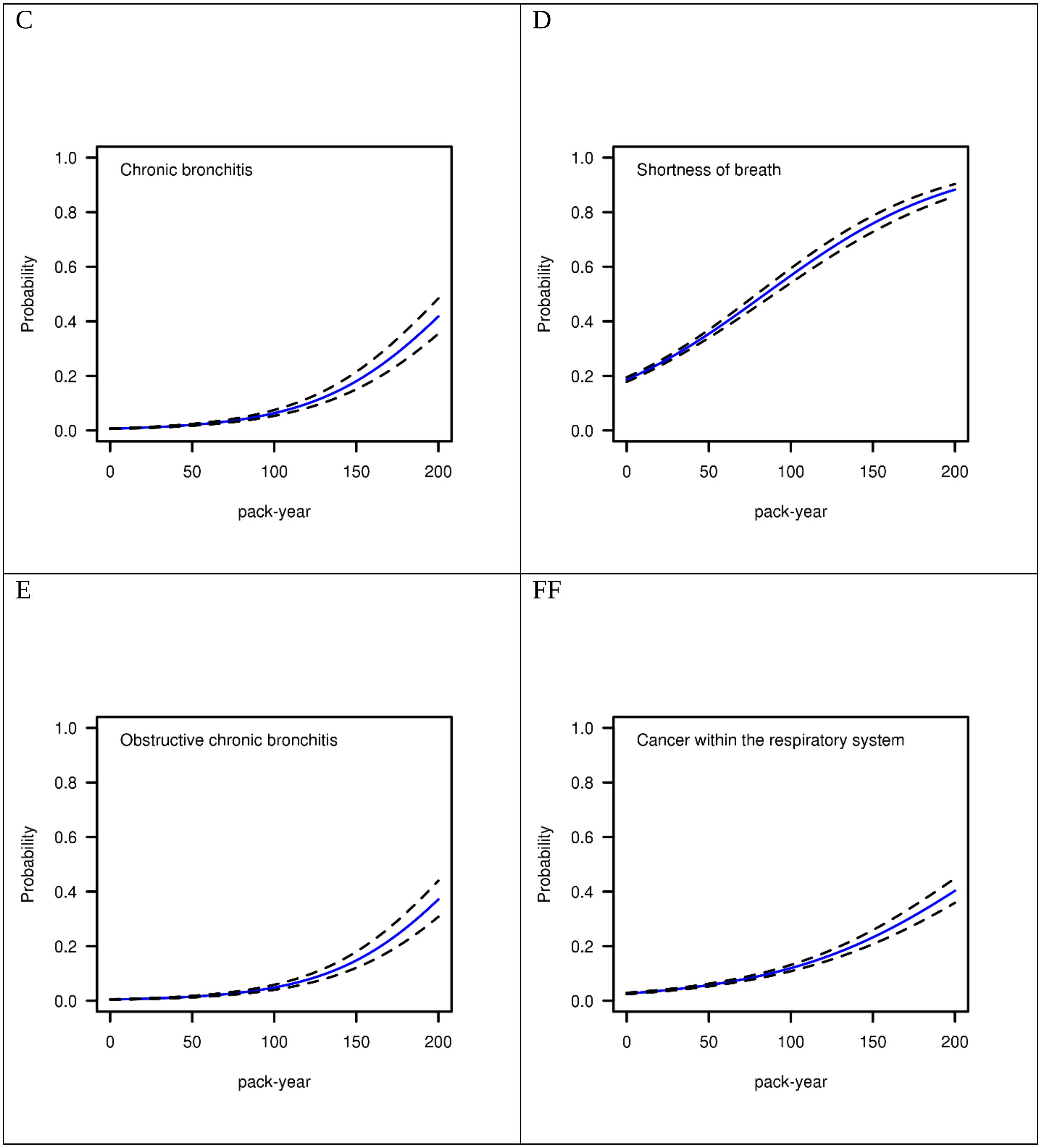
Figure 13: Smoking associations found with SHAPES

The SHAPES PheWAS using pack-year smoking exposure yielded 564 significant associations (Appendix D), Figure 13. The top ten results using SHAPES are shown in Table 15 and Figure 14 shows the probability of having the given phenotype for a theoretical 50 year old man with varying smoking history ranging from 0-200 pack-years.

Description	Phecode	Group	p-value
Chronic airway obstruction	496	Respiratory	<1.0E-305
Tobacco use disorder	318	mental disorders	<1.0E-305
Chronic bronchitis	496.2	Respiratory	1.5E-208
Shortness of breath	512.7	Respiratory	3.2E-204
Obstructive chronic bronchitis	496.21	Respiratory	7.3E-200
Cancer within the respiratory system	165	Neoplasms	7.5E-194
Cancer of bronchus; lung	165.1	Neoplasms	3.8E-192
Emphysema	496.1	Respiratory	2.9E-186
Other dyspnea	512.9	Respiratory	1.3E-175
Pulmonary collapse; interstitial and compensatory emphysema	508	Respiratory	2.4E-174

Table 15. Top 10 smoking-phenotype associations by p-value using SHAPES pack-years PheWAS




C
Chronic bronchitis
D
Shortness of breath
E
Obstructive chronic bronchitis
FF
Cancer within the respiratory system
Probability
1.0
0.8
0.6
0.4
0.2
0.0
pack-year
0
50
100
150
200

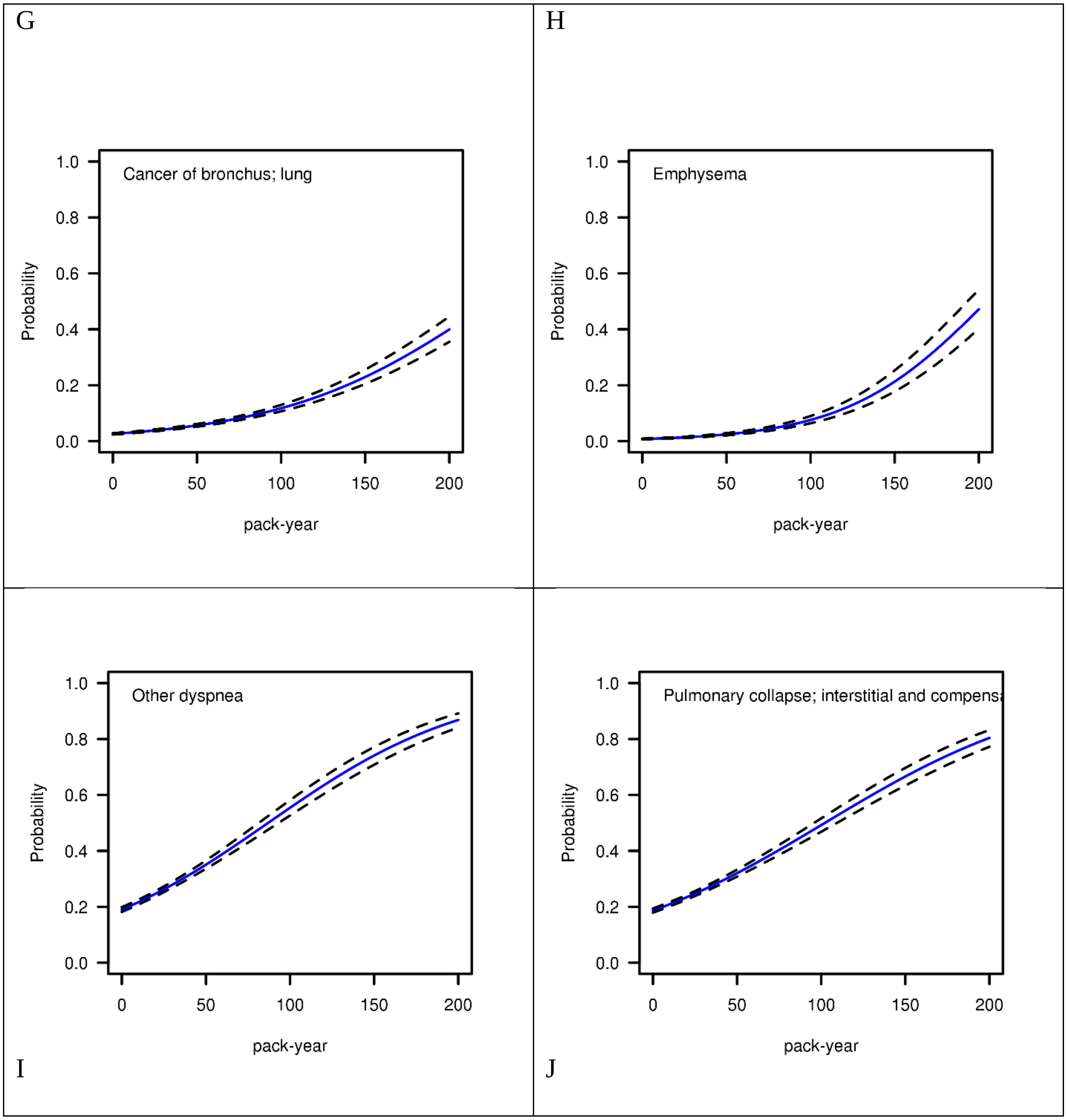


Figure 14. Probability curves generated for a 50 year old man for A) chronic airway obstruction B) documentation of tobacco disorder C) chronic bronchitis D) shortness of breath E) obstructive chronic bronchitis, F) lung cancer (cancer within the respiratory system) G) lung cancer (cancer of bronchus; lung) H) emphysema I) other dyspnea J) pulmonary collapse.

For all but one association, prostate cancer, smoker was a risk factor. Figure 15 shows the relationship between smoking probability of prostate cancer.

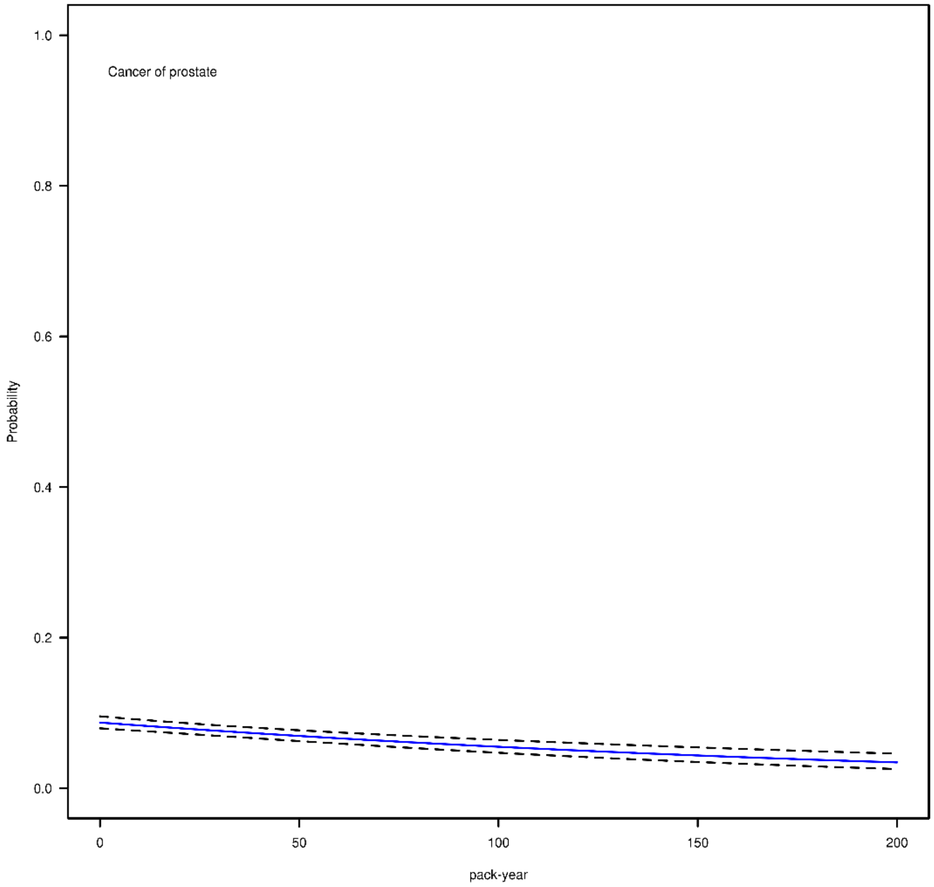


Figure 15. Prostate cancer probability curve

The two methods agreed on 146 of the significant results. Seven results were significant only with the ever-never classification system. SHAPES identified 417 significant results not found with the ever-never classification system. Figure 16 shows the p-value distribution across phenotypes between the two different methods for ascertaining smoking history.

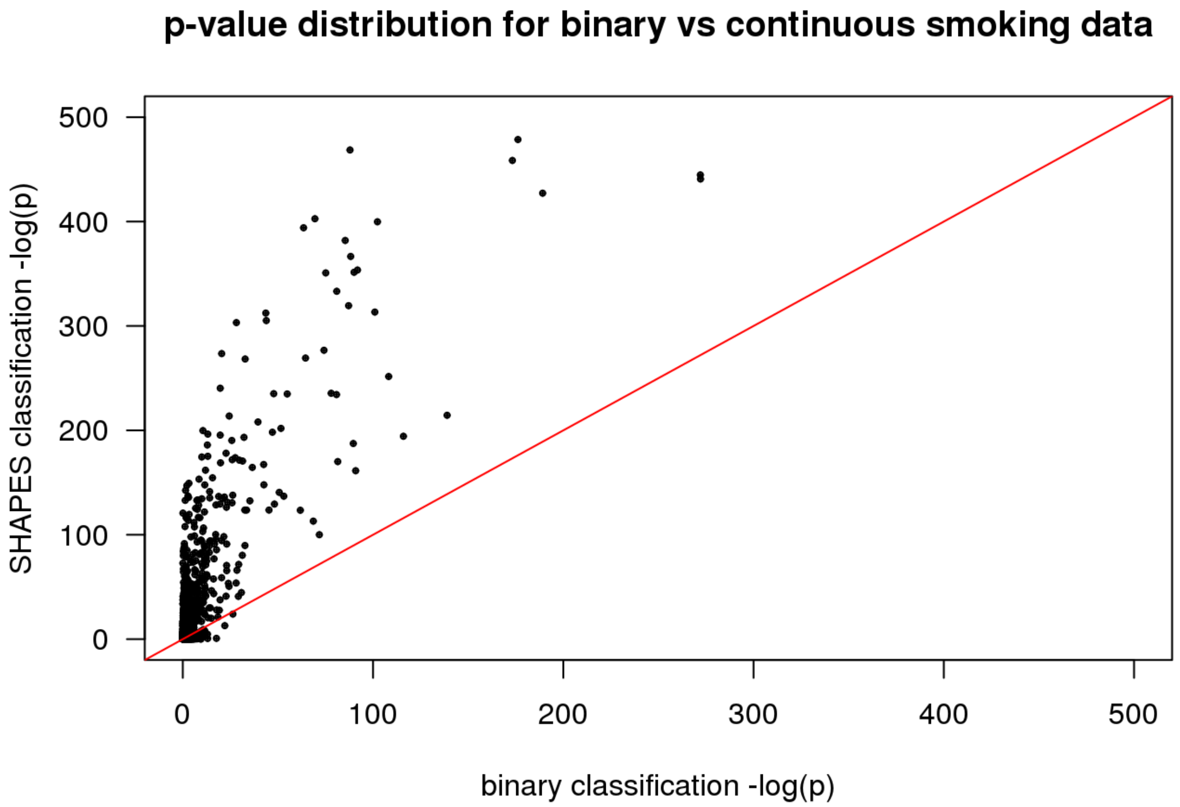


Figure 16. PheWAS p-value comparison between two methods of ascertaining smoking information.

A total of 1574 simulations were performed incrementally removing individuals from the SHAPES analysis and testing SHAPES PheWAS p-value distribution D-statistic against the ever-never p-value distribution. Figure 17 shows the D-statistic plotted as a function of sample size (lower equals closer proximity between the two sets of values.

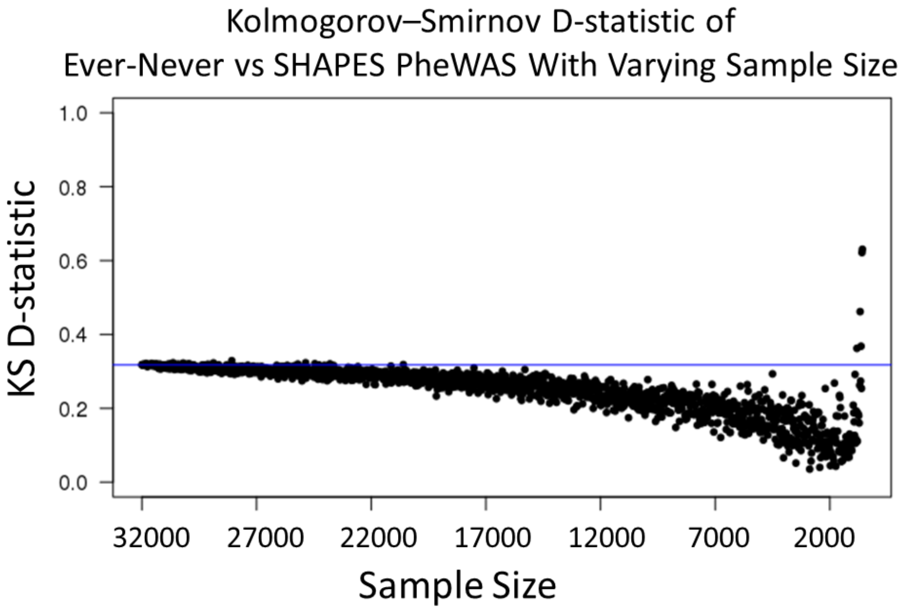


Figure 17. Kolmogorov–Smirnov D-statistic for ever-never smoking PheWAS p-value distribution vs. SHAPES PheWAS p-value distribution as a function of SHAPES PheWAS sample size.

The minimum KS statistic of 0.035 occurred at a sample size of 2368. For illustration, a Manhattan plot is included, Figure 18, showing a PheWAS performed using SHAPES with just 10,000 individuals randomly selected from the prior analysis.

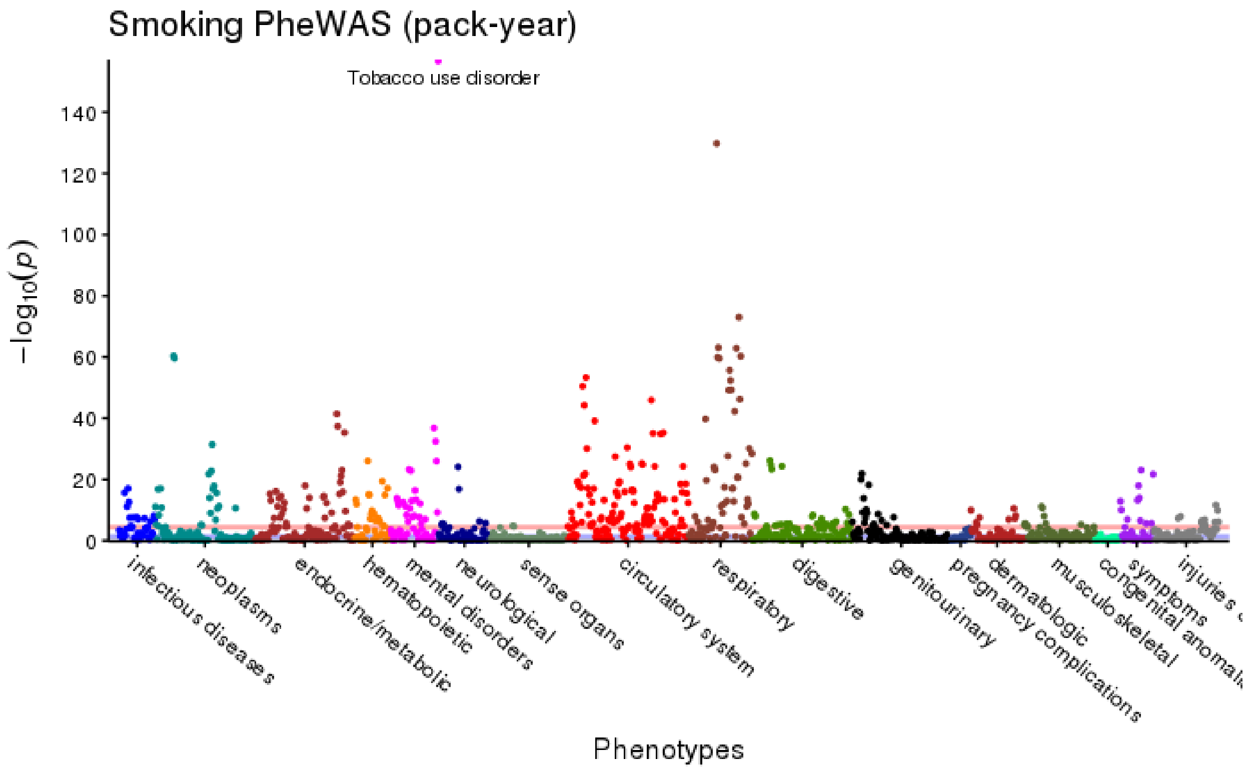


Figure 18. PheWAS of 10,000 individuals randomly selected with pack-years extracted using SHAPES.

Figure 19 shows a p-value plot similar to Figure 16 but comparing 10,000 patients with quantitative smoking data extracted using SHAPES vs 35,788 patients with binary smoking status extracted from the medical record.

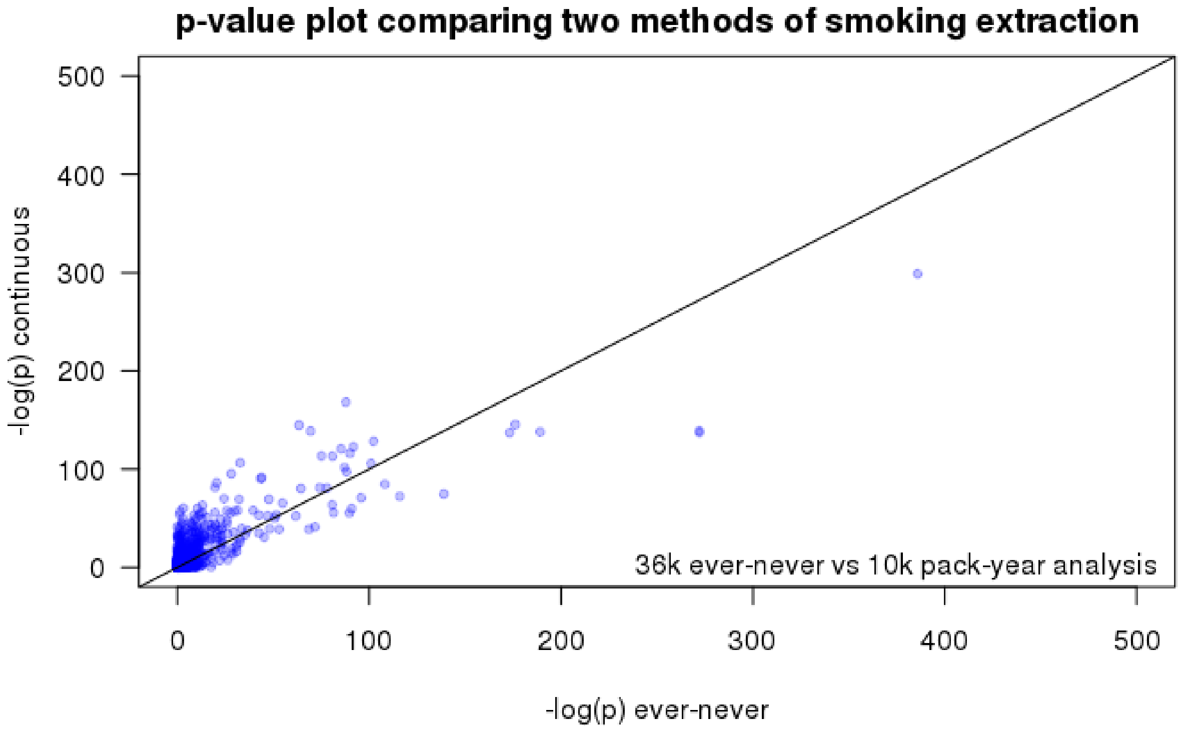


Figure 19. P-value plot comparing p-value distribution for 35,788 patients with smoking binary status extracted and 10,000 patients randomly selected from that group with pack-year tobacco exposure extracted.

Discussion

This analysis was performed on a convenient set selected as they have a higher interaction with the healthcare system (average of 238 billing codes per individual). This may enrich diagnoses and associations. In the two set of individuals, SHAPES was able to identify 5,076 more smoking individuals while the ever-never smoking classification system identified 852 more never-smokers. The number of smoking-disease associations found with the ever-never smoking classifier was not surprising as tobacco has been implicated as a risk factor in numerous diseases. The SHAPES system found over 500 significant associations with 417 results that were not identified with the ever-never classification.

One interesting replication found by both methods was the decreased risk of prostate cancer in individuals who smoke, Figure 14. One proposed mechanism is based on increased rate of circulating sex hormone binding globulin.[85]

Unfortunately, it is difficult to compare odds ratios between the two groups. Table 13 and Appendix C list odds ratios for the ever-never classification system while Figure 14 shows the probability curve based on pack-years. P-values are lower for the majority of significant results (Figure 16) which may be due to similar numbers of smokers found with each method, but SHAPES providing smoking quantity as a continuous variable or SHAPES may be finding false relationships due to inaccurate smoking ascertainment. To help determine which situation is more likely, we simulated PheWAS using incrementally smaller data sets. The Kolmogorov–Smirnov D-statistic provides a method of measuring the distance between the p-value distributions of the SHAPES PheWAS and ever-never PheWAS. Figure 17 shows that as the SHAPES sample size decreases, the D-statistic approaches 0. This implies that the p-values become closer as SHAPES is run on a smaller dataset. This is visualized in Figures 18 and 19 where the same PheWAS is run with 35,788 individuals ever-never smoking information and 10,000 randomly selected individuals from that set are then analyzed using quantitation smoking information provided by SHAPES. With one third of the sample size, the SHAPES PheWAS is able to provide similar results.

Conclusion

This study compares two approaches to ascertain smoking data from the VUMC SD to perform PheWAS. The large number of significant disease associations with smoking is not surprising. We were able to replicate previous results showing a protective effect of smoking on prostate cancer incidence. We also showed that SHAPES was able to find similar associations to commonly-used never-ever smoking classifiers with a much smaller sample size.

CHAPTER IV

A SMOKING GENOME BY ENVIRONMENT (GXE) INTERACTION STUDY

Introduction

Genome by environment (GxE) interaction studies attempt to identify relations between genetic and environment risk for disease. Large-scale GxE interaction studies are traditionally difficulty for two reasons: 1) Interactions are difficult to discover statistically and thus can require large datasets. 2) Genomic and environmental data are not often found in the same system so it is often not possible to integrate the two. The Vanderbilt University Medical Center (VUMC) provides a unique environment to conduct this research as genetic information is provided via the BioVU[50] repository and quantitative smoking exposure is provided via the Smoking History And Pack-year Extraction System (SHAPES, Chapter 2). To our knowledge this is the first phenome-wide GxE study performed on this scale.

Methods

Patients were identified from VUMC's BioVU, a de-identified DNA biobank linked to de-identified electronic medical record data[5]. Patients are given the opportunity to consent and participate in the BioVU biobank during routine care at VUMC. For participants, extra blood remaining after clinical testing can be used for research purposes. Genotype data for this study was conducted by the Vanderbilt DNA Resources Core using the Illumina Human Exome array genotyping platform (details previously reported by Denny, et al).[60] Approximately 36,000 European ancestry adults with Illumina exome array data were selected for this analysis. Smoking status (ever-smoker vs never-smoker) was ascertained on individuals using a validated natural language processing (NLP) and machine learning based system that has been previously described by Liu and colleauges.[44] Quantitative smoking exposures in pack-years were ascertained using SHAPES, Chapter 2. Phenotype or disease statuses were based on PheCodes which are derived from International Classification of Disease Version 9 (ICD9) codes.

We tested 1750 SNP-phenotype pairs which were previously reported in the European Bioinformatics Institute (EMBL-EBI) GWAS catalog significantly associated with a disease and available on the VUMC Illumina exome SNP chip.[86] We attempted to replicate these associations with the entire dataset using a logistic regression with age and gender as covariates (Figure 20, A). We then stratified the population based on smoking status and re-ran the analysis on the two groups (Figure 20, B). Finally, to test for smoking x SNP interaction, we modeled the data using logistic regression with age, gender, smoking status, SNP, and smoking x SNP terms (Figure 20, C)

$$(A)\ phenotype = \alpha + \beta_1 A + \beta_2 Sex + \beta_3 Gen$$

$$(B)\ phenotype = \alpha + \beta_1 A + \beta_2 Sex + \beta_3 Gen + \beta_4 Sm_{bin} + \beta_5 (Gen\ x\ Sm_{bin})$$

$$(C)\ phenotype = \alpha + \beta_1 A + \beta_2 Sex + \beta_3 Gen + \beta_4 Sm_{cont} + \beta_5 (Gen\ x\ Sm_{cont})$$

Figure 20: Logistic regression models for the planned GxE analysis A) replication-only with no smoking or interaction terms B) ever-never smoking status and interaction term C) SHAPES pack-year smoking quantity and interaction term. A: age, Sm_{bin}: ever/never smoking status, Sm_{cont}: smoking quantity in pack-years, Gen: SNP,

P-values were determined based on the genetic term beta value (Figure 20, A) or the interaction term beta (Figure 20, B and C). The significance threshold is reported α less than or equal to 0.05. We also used a two degree of freedom (2df) joint test of the SNP and the interaction term, controlling for the main effects of the remaining covariates.

Results

A total of 31,544 individuals were included in the analysis with ages ranging from less than 1 year of age to greater than 89. The Illumina platform exome SNP chip contained 1588 SNPs on 652 genes reported for this population. A total of 1610 phenotypes were mapped to the study individuals as described above. At the time of analysis, the EMBL-EBI catalog contained 2751 SNP-phenotype pairs. A total of 1750 SNP-phenotype pairs were used as the basis for analysis as those were present in the EMBL-EBI catalog, present on the exome SNP chip, and had phenotype data available for the study population. Only individuals with quantifiable smoking exposure were included in the analysis. Those data were available for 18,830 (59.7%) individuals which contained 15,362 (48.7%) never smokers and 3,468 smokers (11.0%) with greater than zero pack-years of smoking history. The final analysis thus included 18,830 individuals with quantifiable smoking data and 1750 SNP-phenotypes.

The first analysis (Figure 20, A), ignored smoking status and quantity. Of the 1750 SNP-phenotype associations, all of which had previously shown to be statically significant associations in prior studies, 294 (17%) were replicated in this analysis (Appendix E).

Since smoking-phenotype associations have been previously described in Chapter 3, those results will not be repeated here. 57 significant interactions were found between SNP and phenotypic expression (Appendix F). The top 10 results by p-value are included in Table 16.

SNP	PheCode	Description	SNP OR	SNP P-value	Smoking P-value	Interaction P-value
rs10484561	202.21	Nodular lymphoma	1.49	2.4E-05	4.5E-03	4.1E-05
rs2621416	202.21	Nodular lymphoma	1.20	1.9E-03	8.7E-03	9.8E-04
rs1000113	555.1	Regional enteritis	1.43	3.7E-02	1.9E-01	1.7E-02
rs3024505	555.1	Regional enteritis	1.05	7.1E-03	1.1E-01	8.2E-03
rs11747270	555.1	Regional enteritis	1.33	6.7E-02	1.9E-01	1.9E-02
rs7714584	555.1	Regional enteritis	1.33	6.7E-02	1.9E-01	1.9E-02
rs13361189	555.1	Regional enteritis	1.37	6.5E-02	2.0E-01	2.2E-02
rs4846914	272.12	Hyperglyceridemia	1.26	1.4E-04	2.9E-02	1.5E-03
rs2144300	272.12	Hyperglyceridemia	1.26	1.5E-04	3.0E-02	1.6E-03
rs11101442	695.42	Systemic lupus erythematosus	1.00	5.1E-02	5.2E-02	8.7E-03

Table 16: Top ten interactions between SNP and smoking exposure

Evidence of interaction was seen in several cancers, including cancer of the lung, breast, prostate, bladder, liver, and brain. There were also three cardiovascular phenotypes that demonstrated interaction: Ischemic heart disease, dilated cardiomyopathy, and aortic aneurysm; as well as type 1 and 2 diabetes, lupus, rheumatoid arthritis, hypothyroidism and IBD. Of 25 SNP-lung cancer associations that were available for testing in the sample population, six were replicated. All lung cancer-SNP pairs show a strong association with smoking, as expected. Three SNPs (rs1926203, rs7626795, and rs8042374) showed an increased risk of lung cancer only in the individuals with tobacco exposure. Five SNPs showed interaction with tobacco exposure (p <0.05), Table 16.

SNP	Gene	Genetic Risk		Tobacco Risk		Gene-Environment Interaction
		OR	P value	OR	P value	P value
rs7626795	IL1RAP	1.10	0.24	13	6.7E-163	1.3E-03
rs3117582	BAG6	1.08	0.40	11.08	3.2E-140	4.2E-03
rs16951095	LAMA1	1.07	0.48	11.92	1.2E-169	0.01
rs402710	CLPTM1L	1.16	0.01	11.64	3.9E-108	0.01
rs1209950	ETS2	1.02	0.76	12.68	7.1E-94	0.03

Table 17: Significant interactions between lung cancer and SNPs.

Figure 21 shows the relationship between ischemic heart disease and SNP rs17465637 (MIA3 gene on chromosome 1q41). In this analysis 2515 never smokers were compared to 1,081 smokers. Non-smokers did not see an increase in risk of coronary artery disease (CAD) with either heterozygous or homozygous rs17465637 (OR 0.97, p = 0.44). In patients that smoked, however, the risk of developing CAD was 1.25 fold higher ($p = 4.8 \times 10^{-4}$).

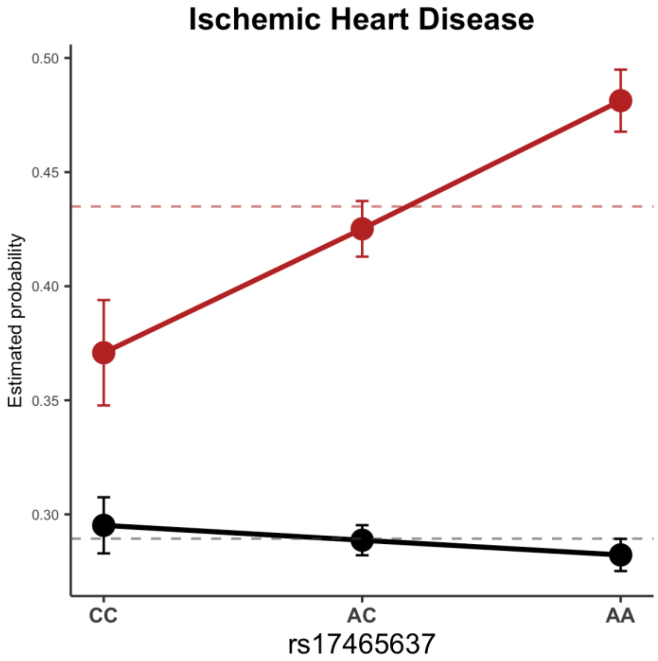


Figure 21: Risk of ischemic heart disease in patients with rs1746537 for non-smokers (black) and smokers (red).

Figure 22 shows the association between SNP rs10871777 and obesity. For never-smokers (n=1,281), rs10871777 did not appear to increase risk of obesity compared to ever-smokers (n=365). The odds ratio for smokers was 1.42 (p=1.5 10^{-4}) compared to 1.04 (p=0.42).

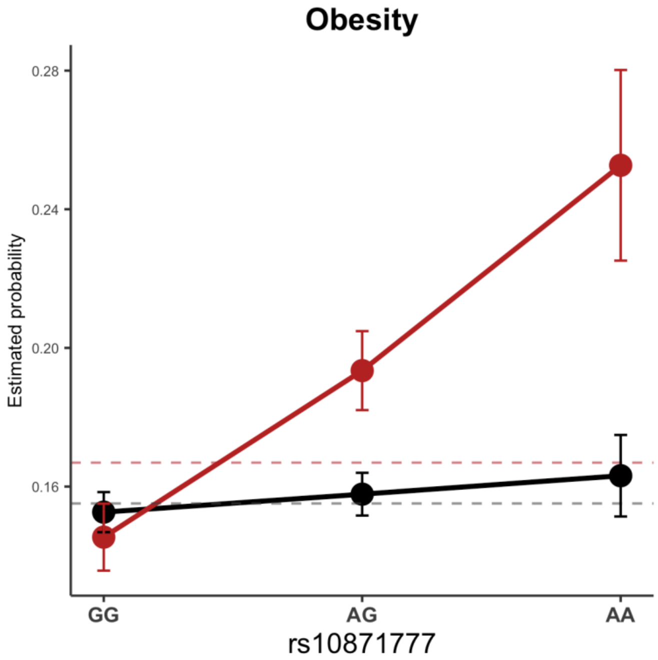


Figure 22: Risk of obesity with smoking and rs10871777 for non-smokers (black) and smokers (red).

Figure 23 shows the risk of type 2 diabetes mellitus (DM) for smokers and non-smokers with or without rs2943641. Without controlling for smoking 3,046 individuals with appear to have a reduced risk of obesity (OR = 0.89, p = 3.1 x 10^{-4}). When controlling for smoking, never-smokers appear to have be the subgroup with the benefit (OR = 0.85, p = 7.2 x10^{-6}), Figure 23. Any benefit from rs2943641 appears to be lost in the presence of smoking (type 2 DM OR = 1.04, p = 0.53).

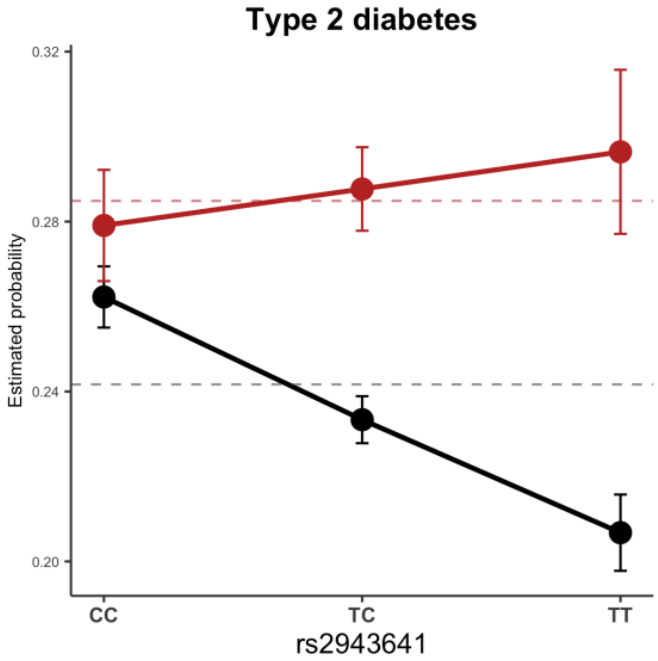


Figure 23: Risk of type 2 diabetes with smoking and rs2943641 for non-smokers (black) and smokers (red).

Discussion

Here we describe a phenome-wide gene by environment interaction study using SNP data from the VUMC BioVU and tobacco exposure extracted using a previously validated ever-never smoking status classifier and SHAPES to extract quantitative smoking data, Chapter 2. The first analysis, a replication study (Figure 20, regression A), showed that 17% of previously discovered SNP-phenotype associations were able to be replicated by this data set when smoking data were ignored. This analysis was not powered to replicate the majority of findings and underlying the importance of having a sufficiently large sample size.

When including smoking data (Figure 20, regression B and C) and testing for interaction, 57 statistically significant interactions between genetic and smoking risk of disease were identified. Since interactions are difficult to discover, it is challenging to assess how many of these nominally significant interaction results are true. We chose to present four cases to demonstrate the utility of this type of analysis 1) lung cancer sub-group 2) ischemic heart disease and rs17465637 3) obesity and rs1081777 and 4) type 2 diabetes and rs2943641.

As shown in Chapter 3, lung cancer is highly associated with smoking and risk depends on the amount of tobacco exposure. Twenty-five SNPs available in the sample population were previously reported to increase the risk of lung cancer. For these SNPs, the risk attributable to genetic factors is dwarfed by the strong association between smoking and lung cancer. Based on this difference in phenotypic expression between genetic and environmental factors, discovering an interaction between seemed unlikely. The rate of significant interaction (5 of 25) was similar to other phenotypes. Three SNPs (rs1926203, rs7626795, and rs8042374) showed an increased risk of lung cancer only in individuals with tobacco exposure. These data may help inform future lung cancer risk calculators and could help refine which patients would most benefit from lung cancer screening by consider environmental and genetic factors.

The SNP rs17465637 resides in the melanoma inhibitory activity family member 3 (MIA3) gene and was shown in a 2007 GWAS to have an association with ischemic heart disease.[87] The increased risk, however, appears to be limited to patients with tobacco exposure. This is collaborated by a 2013 study[88] in which 34% of the population were smokers and a 2011 study[89] where the effect was only replicated after controlling for smoking.

Melanocortin 4 receptor (MC4R) is where rs1081777 is located. A 2015 study investigating the interaction between obesity and adolescent BMI reported interaction between smoking and another MC4R SNP in European-descended adolescents, rs2112347. [90] That study tested 40 interactions between SNPs and smoking for the phenotype of obesity. Two of forty were significant. Figure 22 shows the increase probability of obesity to increase only in smokers with rs1081777. Non-smokers do not seem to carry the same risk, even when homozygous for rs1081777. The other SNP in the Young et al. study, rs1514175, appears to be more common in Hispanic-descended individuals and neither the SNP nor gene, TNNI3K, were significant in our results.

The interaction is not always activated risk with smoking. Figure 23 shows risk for type 2 diabetes with rs2943641 and tobacco. The SNP appears protective in non-smokers, but protective can be overcome in the presence of tobacco exposure. The SNP rs2943641 which is located on the insulin receptor substrate 1 (IRS1) gene, had been initially showed in GWAS studies to be a risk for type 2 diabetes. IRS1 variants affect the rate of insulin resistance and therefore change an individual's risk of type 2 diabetes. A 2013 study by Zheng and colleagues investigates two IRS1 variants, rs2943641 and rs7578326, and concludes that after controlling for a variety of factors, including tobacco use, that IRS1 variants effect insulin resistance, but environmental factors such as dietary intake and modify the association.[91]

Conclusion

This phenome-wide GxE study found 57 statistically significant interactions and reproduced several interactions previously described. Interactions between tobacco exposure and genetic risk were found even in diseases such as lung cancer where the environmental risk is far greater than genetic risk. One main difficulty in studying genome-environment interaction is that a large sample size is required. Ideally this study could be expanded to pool data across multiple healthcare systems. Also, the interpretations of these data are somewhat limited in that the study individuals are nearly all European descent. Additional studies that target a more diverse population and larger samples sizes are warranted.

CHAPTER V

SUMMARY

In this series of studies, we have shown that the Smoking History And Pack-year Extraction System (SHAPES) can extract quantitative smoking history from the electronic health record (EHR). The F-measures of SHAPES at the note and patient-level for classifying smoking status are less than those reported with other systems on other dataset; however, a direct comparison has not yet been performed. The focus of SHAPES is quantitative smoking data extraction, not classification. The main difficulty is with SHAPES' sensitivity and may be addressed through further expert review and rule curation. Despite these limitations, when applied to pragmatic problems such as identifying patients for lung cancer or abdominal aortic aneurysm screening or conducting a PheWAS of smoking using VUMC SD data, SHAPES performed well. In the PheWAS comparison, the system was able to predict similar significant associations with 66% less sample size, and detected 411 (268%) more associations in the full dataset than when using just ever/never status. SHAPES provides a continuous measure extracted from the SD with more precision than recall which is a good use case for studies that are underpowered due to low sample size so could benefit from the higher resolution pack-year variable as opposed to ever-smoker.

Determining genome by environment interactions is then the perfect underpowered use case. Obtaining more genetic data and collecting more tobacco exposure data are challenging. In addition, when compared to other association studies such as GWAS or PheWAS, discovering interactions requires either a larger sample sizes or better variable resolution such as SHAPES provides.

Conclusions

SHAPES provides quantitative smoking data from clinical notes including rate of smoking, duration of smoking, pack-year smoked, and time quit (if applicable). The utility of these measures depends on the research question but is best suited for studies for which a full manual review of smoking data is not possible and that may be under powered using existing smoking status classifiers.

APPENDIX A

Extraction Rules (rules.py)

```
################################################################################
#
# Rules
# ex:
#  from rules import RULES
#  print RULES.CLEAN
#  print RULES.WINDOW.EXCLUSION
#
################################################################################


#
# These are components that can be inserted into expressions below in attempt to
# improve readability
#
TOB                                                                                     =
'(?:(?<!hepa)(?<!pos)(?<!oc)(?<!acine)(?<!lac)tob(?!er|acter|acillus|structive|iliary|
ra|i\s)[a-z]*)'
CIG = '(?<!electronic\s)(?<!e\-)cig[a-z]*'
SMK = '(?:smok(?!ers|eless|e?y|e detector|e exposure)[a-z]*)'
PK      =      '(?:(?<!z      )(?<!z-)(?<!z)(?<!dose-)(?<!dose        )(?<!sone\s)(?<!ice
)(?<!ice)(?<!titration             )(?<!cold             )(?<!cold-)(?<!cold)(?<!flavor
)(?<!tri)(?<!c)p(?:ac)?k(?:age|gs)?s?)(?=[-\s.,/])(?!of|(?:(?:a            )?(?:surgical
)?)?wound|with)'
PPD = '(?:(?<!read )(?<!pos )(?<!neg )(?<!every )(?<!last )(?<!recent )(?<!tuberculin
)(?<!c)(?<!recent     )(?<!\()[-\d\s\./]pp[-dw\s\.,](?!\s(?:test|read|pos|neg|every|[a-
z]{3,12}ly|skin)))'
TIME='(?:y(?:ea)?rs?|mo(?:n|nth)?s?|w(?:ee)?ks?)'
NEVER_MATCH='(?!)' # @see http://stackoverflow.com/questions/1723182/a-regex-that-will-
never-be-matched-by-anything

class RULES:

  LOCAL = 'VUMC'
  VERSION = '0.0.1'

  #
  # replace column 1 with column 2
  # these are local-specific rules and will be the first applied to the document
  #
  CLEAN = [
      [r'\*\*(?:PLACE|INSTITUTION)',r'town'],
      [r'\*\*DATE\[(.*?)\]',r'\1'],
      [r'\*\*NAME\[(.*?)\]',r'\1'],
      [r'\*\*AGE\[birth-12\]',r'12'],
      [r'\*\*AGE\[in teens\]',r'13'],
      [r'\*\*AGE\[in (.*?)s\]',r'\1']
    ]

  BRACKETS = "[\(\)]"

  #
  # determining context window
  #
  class WINDOW:
    INCLUSION = (
      TOB +
      '|'+PK+'[ \-/](?:year|yr)' +
      '|'+PK+'[\s*\-]year' +
      '|'+PPD +
      '|'+SMK +
      '|'+CIG +
      '|'+PK  +
      '|\d\s*py(?!elo)r?[-\s.,]'
    )

    #
    # @TODO what to do with no exclusions?
    #
    EXCLUSION = NEVER_MATCH

  #
  # never smoker
  #
  class NEVER_SMOKER:

    #
    # inclusion rules here have no expectation of group matching
    #
    INCLUSION = [
      '(?:never|no|does(            not|nt)|denie[sd]|neg[a-z]*(?:            |for)*)(?:
|significant|any|ever)*(  |the|us[a-z]+)*(  |of|history  of|h/o|ho|past|habits?)*[  :-
]+('+SMK+'|'+CIG+'|'+TOB+')(?!(in|the\s)*(household|home))',
      '(?:'+TOB+'|'+SMK+'|'+CIG+')(           |significant|any)*(?:           |use)*(?:
|of|history|h/o|ho|past)*[   :-]*(never|no(   history)*|does   not|denie[sd]|none|(is|
)*absent)(?!(in|the\s)*(household|home))',
      '(?:(?:n[a-z]*n|never)[                                                          -
]*('+TOB+'|'+SMK+'|'+CIG+'))\s(?!(in|the)*(household|home))',
      '(?:never|no|does( not|nt)|denie[sd]|neg[a-z]*(?: |for)*)(?: |significant|any)*(
|use)*(                     |of|history                        of|h/o|ho|hx|past)*[
:]*((alcohol|drink|etoh|drug[s]?|abuse|illicit|recreational|chew[a-
z]*|cocaine|any|use)*(,|
|and|or|either)+)+('+TOB+'|'+SMK+'|'+CIG+')(?!(in|the\s)*(household|home))',
      '(?:lifetime )?(?:non|never).?'+SMK,
      'non?.?'+TOB,
      'does not(?:[\s,]|consume|drink|alcohol|or)+('+TOB+'|'+SMK+')',

'not?(?:[\s,]|have|or|use|diabetes|caffeine|cocaine|alcohol|ethanol|illicit|drug)+('+C
IG+'|'+TOB+')',
      '(?:w/o|without|no) (?:a )?(?:hx|history)(?: of)? (?:' + TOB + '|' + SMK + '|' +
CIG + ')',
    ]

    #
    # exclusion rules here have no expectation of group matching
    #
    EXCLUSION = [
      '(?:remote[a-z]*|distance|past|former[a-z]*|decrease|no      longer|\s){2,10}(?:
|in|of|the|amount|total|'+CIG+'|pipe)*(?:'+TOB+'|'+SMK+'|'+CIG+')',
#              '(?<!no  )(?:'+TOB+'|smok[a-z]*(  of)?|'+CIG+')+(  |in|the)*(remote[a-
z]*|distance|past|former[a-z]*|decrease|no longer|in|the|(?<!no )history| )+',
      'year ('+TOB+'|'+SMK+'|'+CIG+') history',
      SMK+' cessation',
      '(?:quit|amount  of|back  to|stopped)[  ]*('+SMK+')*[  ]*(.*year[s]?|ago|age|in|
|\d{4}){2,10}',
      '(?:not|no) ('+SMK+'|'+TOB+CIG+') (now|recent[a-z]*|lately|today|this .*)',
      'ever '+SMK+': yes',
      '(?:no|has
not|hasnt)\s+('+SMK+'|'+CIG+'|'+TOB+')\s+(in|since|until|after|before)\s\d',
      TOB+' use disorder',
      SMK+'\s(x|times|for)',
      '(?:former|prior|used to) '+SMK,
      '(?<!no )(?:'+SMK+' :)$',
      '(?:'+TOB+'|'+SMK+'|'+CIG+')(?:in|the|\s)*(?:home|household)',

'(?:dad|mom|sister|brother|uncle|aunt|father|mother|grandfather|grandmother)\s'+SMK,
      '(?:quit|amount of|back to|stopped)[ ]*('+SMK+')*[ ]*(.*year[s]?|ago|age|in|
|\d{4}){2,10}',
      '(?:not|no) ('+SMK+'|'+TOB+'|'+CIG+') (now|recent[a-z]*|lately|today|this .*)',
      '(?:no|has
not|hasnt)\s+('+TOB+'|'+SMK+'|'+CIG+')\s+(in|since|until|after|before)',
      '(?:former|prior|used to|off)\s(?:'+SMK+'|'+TOB+'|'+CIG+')',
      'stopped '+SMK,
      '(?:' + SMK + ') (?:\s|some|in|the|very|distant|former[a-z]*|remot[a-z]*)+past',
      'quit\s'+SMK,
      '(?<!any )(?<!no )(?:history|hx)\sof\s'+SMK,
    ]

  #
  # ever smoker
  #
  class EVER_SMOKER:

    #
    # inclusion rules here have no expectation of group matching
    #
    INCLUSION = [
        '(?:(dis)?continue[sd])?\s*()?\s*('+SMK+')',
        PPD,
        '(?:'+CIG+'|'+PK+')',
        '(?:per|a|every|q)\s*[\d+\-
to\s]*\s*(d|days?|weeks?|wks?)?|daily|weekly|montly',
        '(?:'+SMK+')',
        '(?:quit|amount of|back to|stopped)[ ]*('+SMK+')*[ ]*(.*year[s]?|ago|age|in|
|\d{4}){2,10}',
        '(?:not|no) ('+SMK+'|'+TOB+'|'+CIG+') (now|recent[a-z]*|lately|today|this .*)',
        '(?:no|has
not|hasnt)\s+('+TOB+'|'+SMK+'|'+CIG+')\s+(in|since|until|after|before)',
        '(?:former|prior|used to) '+SMK,
        'stopped '+SMK,
        '(?:' + SMK + ') (?:\s|some|in|the|very|distant|former[a-z]*|remot[a-
z]*)+past',
      ]

    #
    # exclusion rules here have no expectation of group matching
    #
    EXCLUSION = [
         '(?:adults?|mother|father|
son|daughter|niece|nephew|mom|dad|grandfather|wife|husband|girlfriend)[\s-
]*(?:was|is|uses|died|with|a| )*\s*'+SMK,
         'family history of [-\w\s]+(?:,[-\w\s]*)*('+TOB+'|'+CIG+'|'+SMK+')',
         'No '+PK+' per day',
         '(?:discuss[a-z]*|attend[a-z]*)\s*(avoid[a-
z]*|class|session)(\s|of|on)*('+SMK+'|tob)',
         '(?:echo[a-z]*|study)[\w *]*smoke',
         SMK+'(?:\s|in|the)+(?:house(?:hold)?|home)',
         '(?:father|mother|parents|siblings?|sister|brother)\s'+SMK,
      ]
  #
  # duration
  #
  class DURATION:
    PRE = '(?<!quit )(?<!stopped )(?<!to stop )'
    _d = '[~]?(\d+(?:\.\d+)?)'

    #
    # order by specificity
    # inclusion rules here have expect single group match
    # use (?:) for all non-match groupings
    #
    INCLUSION = [
      '('+PRE+'(?:(?:'+PPD+'|(?:'+PK+'(?:  of)?(?:  '+CIG+')?|'+CIG+')[\s]*(?:(?:a|per)
day|daily)|'+SMK+'))[\s\.]*(?:x|times|for|\>|more|last|over|on|and|the|off|than|only|a
t|least|some|intermitantly|\s){2,18}'+_d+'\s*'+TIME+'[\s,\.]+(?!(old|f/u|follow[-
\s]up|ago)))',

'('+PRE+'(?:'+PPD+'|'+PK+'\s*(?:per|/)\s*day|'+SMK+')\s(?:x|times|for)[\s]?'+_d+'\s*'+
TIME+'?(?! day))',
      '('+PRE+'(?:'+CIG+'|'+PK+')                                    (?:per|every|each)
(?:day|d|w[e]*k)[\s\.]*(?:for|x|times) '+_d+'[\s\+]*'+TIME+')',
      '('+PRE+SMK+'(?:for|\s)+'+_d+'[\s\+]*'+TIME+')',
      '('+PRE+''+_d+'[\s\+]*('+TIME+'|history|of|\s)+ ('+TOB+'|'+SMK+'))',

'('+PRE+'(?:'+TOB+'|'+SMK+'|in|the|past|for|over|more|than|only|at|least|intermitantly
|'+PPD+'|\s)+[\s\.]*(?:x|times|for)
(?:more|than|only|at|least|about|around|over|nearly|approx[a-
z]*|\s)+'+_d+'\s*'+TIME+'?(?! day))',
      #   '('+PRE+'(?<!no    )'+TOB+'    (?:use|history|hx|includ[a-z]*|    ){0,4}(?:[-
\s\.:x]{1,6}|for)+'+_d+'[\s\+]*'+TIME+')',
      '('+PRE+'(?<!no      )'+TOB+'      (?:use|history|hx|includ[a-z]*|      ){0,4}(?:[-
\s\.:x]{1,6}|for){1,6}'+_d+'[\s\+]*'+TIME+')',
      '('+PRE+'(?:'+PPD+'|'+PK+'                                              (?:per|a)
day|'+SMK+')(?:on|off|and|occasionally|for|x|\s){3,14}'+_d+'\s*'+TIME+'?(?! day))',
      '('+PRE+TOB+' use[\.: ]+x\s'+_d+'[\s\+]*'+TIME+'?)',
      '('+PRE+TOB+'[-: \.]*'+_d+'[\s\+]*'+TIME+' (?:history|hx))',
      '('+PPD+' '+_d+'[\s\+]*'+TIME+')',
      # '('+_d+'\s'+TIME+'(\s|of|'+SMK+'|(?:[\d\.]+\s*ppd)){4,10})',
      '('+_d+'\s'+TIME+'(\s|of|'+SMK+'|(?:[\d\.]+ '+PPD+')){4,10})',
      '(since (?:he |she |they )?(?:was |were )?x '+_d+' years)',
#
'('+PRE+'\s*(?:'+TOB+'|'+PPD+'|'+SMK+'|'+PK+'|of|since|child|a|less|younger|than|age|o
n|and|off|than|only|at|least|some|intermitantly|at|least|about|around|over|nearly)+\s*
(\d\d))'
    ]

    #
    # [DD] represents the digit extracted from the inclusion rule
    # no expectation of group matching
    #
    EXCLUSION = [
      '(in|on) [DD][-\s\.,]',
      # '(quit|not|stop[a-z]*) smok[a-z]* (x|times|for|\s)*[DD][-\s\.,]',
      '(?:quit|stopped|not)                     (?:(?:'+PPD+'|(?:'+PK+'|'+CIG+')[\s]*per
day|'+SMK+')[\s]*(?:x|times|for)|\>|more|over|on|and|off|than|only|at|least|\s){2,18}[
DD]\s*'+TIME+'[\s,\.]+(?!(old|f/u|follow[-\s]up|ago))',
    ]

    RANGE = [
      ['(from (\d{4}) to (\d{4}))',[1,2]],
      ['(from(?:\s|approx[a-
z]*|about|around)+(?:age\s)?(\d{1,2})(?:\s|to|until|approx[a-
z]*|about|around){3,10}(?:age\s)(\d{1,2}))',[1,2]],
      ['(from age\s*(\d{1,2})\s*(?:until|to|-)\s*(?:age)?\s*(\d{1,2}))',[1,2]],
    ]

  class PACKYEAR:

    #
    # explicit quantity
    # inclusion rules here have expect single group match
    # TODO replace with _d
    INCLUSION = [
      '(\d+)(?:-|\s|\+|\.|plus)*(?:'+PK+'[ \-/]*(?:years?|yr)|pyr?)',
      '(\d+)(?:-|\s|\+|\.|plus)*years? '+PK,
      '(\d+)(?:-|\s|\+|\.|plus)*ppys?',
    ]
    # [DD] represents the digit extracted from the inclusion rule
    EXCLUSION = [
      '(?:quit|'+SMK+'|\s){2}[DD](?!\d)'
    ]

  class QUIT:
    HAS_QUIT = [
      '(?<!in )(quit[a-z]*|former) (' +TOB+'|'+SMK+'|'+CIG+')',
      'no[t]? longer '+SMK,
      '(?<!no)(?<!denies)(?:remote[a-z]*|distance|past|former[a-z]*|decrease|no
longer|\s){2,10}( |in|of|the|amount|total|'+CIG+'|pipe)*('+TOB+'|'+SMK+'|cig)',
      '(?<!no)(?<!denies)('+TOB+'|'+SMK+'(      of)?|'+CIG+')+(      |in|the)*(remote[a-
z]*|distance|past|former[a-z]*|decrease|no longer|in|the|history| ){2,10}',
      '(?<!no)(?<!denies)(?:any|remote[a-z]*|distance|past|former[a-z]*|decrease|no
longer|in|the|history|of| ){2,16}'+SMK,
      '(?:quit|amount  of|back  to|stopped)[  ]*('+SMK+')*[  ]*(.*year[s]?|ago|age|in|
|\d{4}){2,10}',
      '(?:not|no) ('+SMK+'|'+TOB+'|'+CIG+') (now|recent[a-z]*|lately|today|this .*)',
      '(?:no|has
not|hasnt)\s+('+TOB+'|'+SMK+'|'+CIG+')\s+(in|since|until|after|before)',
      '(?:former|prior|used to) '+SMK,
      # way worse '(?:former|prior|used to|off)\s(?:'+SMK+'|'+TOB+'|'+CIG+')'
      'stopped '+SMK,
      '(?:'+SMK+'[a-z\d]*|\s)+past',
      '(?:(?:previous|prior)\s(?:'+SMK+'|'+TOB+'|'+CIG+'))',
      ]

    YEARS_AGO = [
      '((?:(?:quit|history(?:   of)?|hx(?:   of)|discontinued?)\s*(?:'+SMK+'|'+TOB+'(?:
use)?)|(?<!to  )stop[a-z]*  '+SMK+')\s*(?:over|at|x|for|greater|least|than|of|approx[a-
z]*|~| )+(\d+) '+TIME+'(?: ago)?)',
      '((?:quit\s*(?:'+SMK+'|'+TOB+'|       )+(?:over|at|x|greater|least|than|approx[a-
z]*|~| )+)(\d+) '+TIME+'(?: ago)?)',
      '((?:'+SMK+'|'+TOB+')(?:up|until|approx[a-z]*|   ){2,10}[   ~]*(\d+)   '+TIME+'(?:
ago)?)',
      '((?:stopped|quit)  (?:'+SMK+'|'+TOB+'|'+CIG+'|\s)+(?:|x|~|about|around|approx[a-
z]*)+(\d+) '+TIME+'(?: ago))',
      '(if patient quit, when was it:[a-z ~]*(\d+)[a-z ]*)',
      '(quit(?: '+SMK+')?(?:[ ~]|about|around|approx[a-z]*)+(\d+) '+TIME+' ago)',
      '((?:stopped|quit)  '+SMK+'  (?:'+CIG+'|x|for|\s|about|around|approx[a-z]*)+(\d+)
'+TIME+')',
      '(' + SMK+'(?:remotely|until|the|mid|late|\s){3,16}x (\d+) '+TIME+')',
      '(' + SMK+'(?: up) until (?:x|in) (\d+) '+TIME+')',
      '(no (?:'+SMK+'|'+CIG+'|'+TOB+')(?:x|in|for|about|\s){3,16}(\d+) '+TIME+')',
      '(smoke[-\s]?free\s(?:over|at|x|for|greater|least|than|of|approx[a-z]*|~|
)+(\d+)\s'+TIME+')',
  ]

  _m                                                                        =
'(?:jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec|january|february|march|april|june|
july|august|september|october|november|december)'
  INCLUSION = [
   '(?:quit '+SMK+'|stop[a-z]*|remotely|'+SMK+'|quit)\s*in(?:the| )+(\d{2,4})',
   '(?:'+SMK+'|'+TOB+')(?: up)?(?:until|approx[a-z]*| )+ (\d{2,4})',
   '(?:quit|stopped)(?:'+SMK+'|'+TOB+'|\s)+(?:in|on|around|\s)*(\d{4})',

'(?:quit|stopped)(?:'+SMK+'|'+TOB+'|\s)+(?:in|on|around|\s)*'+_m+'[\s,\d]+(\d{4})',

'(?:quit|stopped)(?:'+SMK+'|'+TOB+'|\s)+(?:in|on|around|\s)*\d{1,2}/d{1,2}/(\d{2,4})',
   SMK + '(?:up|until|a|year|\s)+(\d{4})',
   '(?:quit|stopped)\s(?:'+TOB + '|' + CIG + '|' + SMK + ')\s*'+_m+'\s*(\d{4})',
  ]


 class RATE:
  _ex = '\+?[-\s]*((of|a| )*('+CIG+'|'+PK+'|of|[-\s]){2,10}((/|per|a|every|other|q|[-
\s])+[\d+\-to]*(d|days?|weeks?|wks?)|daily|weekly|montly)|'+PPD+')[\s\.,=:]'
  _d = '(?<!\d)\d(?:\.\d+)?'
  _f                                                                        =
'(?:(?:/|per|a|every|other|q|\s){2,10}(?:d(?:ay)?s?|weeks?|wks?)|(?:daily|weekly|montl
y|\s){2,10})'

  # order by specificity
  INCLUSION = [
   '((?<!\d)(\d{1,3}(?:\.\d+)?)' + _ex+')',
   '(('+_d+')\s?pp\sq\dd)',
   '(('+_d+')\s?(?:'+PK+'|'+CIG+'|of|/|\s)+every (?:\d+(?:\.\d+)?|other) days?)',
   '(('+_d+')\s?(?:'+PK+'|'+CIG+'|of|\s)+'+_f+')',
   '(('+_d+') '+PK+'/day)',
   '(('+_d+') pp q)',
   '('+TOB+'\s('+_d+')\s'+_f+')',
   '('+PPD+':\s?('+_d+'))',
   '(('+_d+')\s?'+CIG+'(?:/|'+_f+'))',
   #(\d+) cig/day',
  ]


  # [DD] represents the digit extracted from the inclusion rule
  EXCLUSION = [
   '(quit|not|stop[a-z]*) '+SMK+' (x|times|for|\s)*[DD][-\s\.,]',
  ]

  IMPLIED_ONE   =   '(?<!\d)(?:\d{4})?\s(?:'+PK+'   (?:(?:a|per)   day|daily)|(?<!had
)(?<!\s\-\s)(?<!many                                                        )(?<!and
)'+PPD+'(?!:)(?!\sskin|\splaced|\seach|\severy|\sannually|\shas|\sneg|\spos))'
  CIGS = '\s*(cigs?|cigarettes?)'
  PER_WEEK        =        '\d+\+?\s*(('+CIG+'|'+PK+')\s*((/|per|a|every|q)\s*[\d+\-
to\s]*\s*(weeks?|wks?|weely))|ppw)'
  PERIODIC = [
   ['(?:'+PPD+'|packs?(?:   of   '+CIG+')?|'+TOB+')\s?(every   other   day|qod|q   2
d|qow|qom|qoy)',1],
   ['(?:'+PPD+'|packs?(?: of '+CIG+')?|'+TOB+')\s?q\s?(\d+)\s?d',1],
   ['(?:'+PPD+'|packs?(?: of '+CIG+')?|'+TOB+')\severy\s(\d+)\s?(?:days?|d\s)',1],
  ]

 class TIME:
  AGE_EXTRACTION = [
   '((?:(?:starting|at|since|age|of|the|\s+){3,10})[\']?(\d{2}))(?!s|\d)',
  ]
```

```
    YEAR_EXTRACTION = '((?:starting|in|since|the|late|~|\s){3,16}(\d{4})s?)(?!\d)'

    SINCE_AGE = [

['((?:since|from)(?:\s|a|child|less|than|approximately|about|almost|nearly)+\s(\d+)\sy
[ea]*rs\s(?:of age|old))'],

['((?:since|from)(?:\s|he|she|was|has|been|approximately|about|less|than|almost|nearly
)+age\s(\d{1,2}))'],
    ]

    DECADE_PERSON_YY = [

r'(?<![\d\w])((?:\s|was|s?he|since|has|been|approximately|about|less|than|almost|nearl
y)*(?:\s|in (?:his|her|their)(?:mid|late|early)?)+(\d\d)s)',
    ]

    DECADE_YY = [
      r'(?<![\d\w])((?:\s|in(?! his| her| their)|the|mid|late|early)+(\d\d)s)',
    ]

  class REASONABLE:
    LIMIT = {
      'pack years': {
        'max': 400,
        'min':  -1,
      },
      'rate': {
        'max':  6,
        'min': -1,
      },
      'duration': {
        'max':  100,
        'min': -1,
      },
      'years_quit': {
        'max':  100,
        'min': -1,
      },
      'never smoker status': {
        'max':  1,
        'min': -1,
      },
      'never smoker status': {
        'max':  1,
        'min': -1,
      },
      'has quit': {
        'max':  1,
        'min': -1,
      },
    }

  class NUMBER:

    # RANGE = '(\d+\.?[257]*(?:\s*(?:'+PK+'?\s*)(?:\-|to)[ ]?)+(\d+\.?[257]*))'
    RANGE = [
      '(\d+\.?[257]*(?:\s*(?:\-|to)[ ]?)+(\d+\.?[257]*))',
      '(\d+\.?[257]*(?:\s*(?:'+PK+'?\s*)(?:\-|to)[ ]?)+(\d+\.?[257]*))',
    ]

    # most conventional word mappings are taken care of by
    # numeral_extractor.py
```

```
MAP = [
  ['one and (?:a )?half','1.5'],
  ['1[-\s]1/2','1.5'],
  ['2[-\s]1/2','2.5'],
  ['3[-\s]1/2','3.5'],
  ['4[-\s]1/2','4.5'],
  ['5[-\s]1/2','5.5'],
  ['(?:one[ \-])*half','0.5'],
  ['half','0.5'],
  ['1[-\s]1/2','1.5'],
  ['1/2','0.5'],
  ['(?:a|one)[- /]*(?:3|third)','0.333333333333333'],
  ['(?:a|one)\s(?:3|third)','0.333333333333333'],
  ['(?:a|one|1)/(?:3|third)','0.333333333333333'],
  ['(?:two)[- /]*(?:3|third[s]?)','0.666666666666666'],
  ['(?:two)\s(?:3|third[s]?)','0.666666666666666'],
  ['(?:two|2)/(?:3|third[s]?)','0.666666666666666'],
  ['(?:a|one)[- /]*(?:4|fourth|quarter)','0.25'],
  ['(?:a|one)\s(?:4|fourth|quarter)','0.25'],
  ['(?:a|one|1)/(?:4|fourth|quarter)','0.25'],
  ['(?:three)[- /]*(?:4|fourth[s]?|quarter[s]?)','0.75'],
  ['(?:three)\s(?:fourth[s]?|quarter[s]?)','0.75'],
  ['(?:three|3)/(?:4|fourth[s]?|quarter[s]?)','0.75'],
  ['a few',     '3'],
  ['1(?:\s|and|a)+0.5',     '1.5'],
  ['(?:a )?teen',   'age 13'],
  ['(?:a )?teenager', 'age 13'],
  ]
```

APPENDIX B

SHAPES Dependencies

- python 2.7
- pyparsing
- flask (web implementation)
- tqdm
- sklearn
- scipy
- numpy
- pandas
- seaborn
- tabulate
- termcolor (interactive note review)

APPENDIX C

Significant results with ever-never smoking classification system

Description	Phecode	group	P-value	OR
Tobacco use disorder	318	mental disorders	5.2E-286	6.8
Chronic airway obstruction	496	respiratory	2.6E-168	3.6
Cancer of bronchus; lung	165.1	neoplasms	6.6E-119	4.5
Cancer within the respiratory system	165	neoplasms	7.2E-119	4.4
Emphysema	496.1	respiratory	7.19E-83	6.7
Chronic bronchitis	496.2	respiratory	3.12E-77	4.6
Obstructive chronic bronchitis	496.21	respiratory	5.76E-76	5.3
Alcohol-related disorders	317	mental disorders	4.61E-61	5.0
Secondary malignant neoplasm	198	neoplasms	4.38E-51	1.8
Substance addiction and disorders	316	mental disorders	1.01E-47	3.1
Pulmonary collapse; interstitial and compensatory emphysema	508	respiratory	3.71E-45	1.7
Atherosclerosis	440	circulatory system	1.45E-44	2.3
Coronary atherosclerosis	411.4	circulatory system	1.33E-40	1.7
Alcoholism	317.1	mental disorders	3.43E-40	4.7
Ischemic Heart Disease	411	circulatory system	8.06E-40	1.6
Acute pain	338.1	neurological	1.2E-39	1.9
Other diseases of lung	510	respiratory	4.75E-39	1.8
Shortness of breath	512.7	respiratory	6.8E-39	1.6
Myocardial infarction	411.2	circulatory system	1.42E-38	2.0
Respiratory failure, insufficiency, arrest	509	respiratory	8.35E-38	1.7
Atherosclerosis of native arteries of the extremities with intermittent claudication	440.22	circulatory system	4.24E-36	3.9
Pleurisy; pleural effusion	507	respiratory	7.24E-36	1.7
Empyema and pneumothorax	506	respiratory	8.17E-36	2.4
Atherosclerosis of the extremities	440.2	circulatory system	1.35E-34	2.6
Respiratory failure	509.1	respiratory	2.33E-33	1.8
Peripheral vascular disease, unspecified	443.9	circulatory system	5.92E-33	2.2
Secondary malignancy of respiratory organs	198.2	neoplasms	6.78E-32	2.1
Other dyspnea	512.9	respiratory	6.95E-31	1.6
Secondary malignancy of lymph nodes	198.1	neoplasms	1.62E-30	1.8
Peripheral vascular disease	443	circulatory system	9.69E-29	2.0
Other symptoms of respiratory system	512	respiratory	2.68E-28	1.4
Chemotherapy	197	neoplasms	1.5E-27	1.6
Symptoms involving respiratory system and other chest symptoms	519.9	respiratory	1.46E-24	2.0
Cancer of larynx, pharynx, nasal cavities	149	neoplasms	8.78E-24	2.1

Description	Phecode	group	P-value	OR
Mood disorders	296	mental disorders	3.93E-23	1.5
Cancer, suspected or other	195	neoplasms	9.6E-23	1.7
Cancer of larynx	149.4	neoplasms	1.2E-21	2.1
Other diseases of respiratory system, not elsewhere classified	519	respiratory	1.75E-21	1.8
Depression	296.2	mental disorders	3.55E-21	1.5
Abdominal aortic aneurysm	442.11	circulatory system	2.1E-20	2.7
Pneumonia	480	respiratory	8.67E-20	1.4
Nonspecific chest pain	418	circulatory system	1.14E-19	1.3
Other aneurysm	442	circulatory system	3.1E-19	2.0
Chronic obstructive asthma	495.1	respiratory	3.46E-19	3.1
Abnormal findings examination of lungs	514	respiratory	6.86E-18	1.6
Anxiety, phobic and dissociative disorders	300	mental disorders	1.35E-16	1.4
Aortic aneurysm	442.1	circulatory system	4.83E-16	2.0
Lymphadenitis	289.4	hematopoietic	2.75E-15	1.6
Cough	512.8	respiratory	5.84E-15	1.4
Cancer of mouth	145	neoplasms	6.41E-15	2.0
Dependence on respirator [Ventilator] or supplemental oxygen	509.8	respiratory	7.01E-15	2.4
Other chronic ischemic heart disease, unspecified	411.8	circulatory system	1.07E-14	1.6
Cerebrovascular disease	433	circulatory system	2.27E-14	1.4
Bipolar	296.1	mental disorders	2.59E-14	2.2
Viral hepatitis C	70.3	infectious diseases	3.84E-14	2.2
Alcoholic liver damage	317.11	mental disorders	4.22E-14	4.7
Occlusion and stenosis of precerebral arteries	433.1	circulatory system	1.16E-13	1.6
Secondary malignancy of brain/spine	198.5	neoplasms	1.74E-13	1.9
Certain early complications of trauma or procedure	958	injuries & poisonings	2.05E-13	2.4
Cancer of oropharynx	149.1	neoplasms	4.38E-13	2.5
Disorders of fluid, electrolyte, and acid-base balance	276	endocrine/metaboli c	5.81E-13	1.3
Traumatic and surgical subcutaneous emphysema	958.2	injuries & poisonings	6.7E-13	4.0
Congestive heart failure (CHF) NOS	428.1	circulatory system	1.06E-12	1.4
Cancer of esophagus	150	neoplasms	3.67E-12	2.7
Elevated white blood cell count	288.2	hematopoietic	4.01E-12	1.5
Diseases of white blood cells	288	hematopoietic	5.28E-12	1.4
Painful respiration	512.2	respiratory	5.67E-12	2.0
Congestive heart failure; nonhypertensive	428	circulatory system	6.45E-12	1.3
Tachycardia NOS	427.7	circulatory system	2.54E-11	1.4
Cancer of tongue	145.2	neoplasms	2.82E-11	2.3

Description	Phecode	group	P-value	OR
Secondary malignancy of bone	198.6	neoplasms	3.98E-11	1.6
Pulmonary insufficiency or respiratory failure following trauma and surgery	509.3	respiratory	6.4E-11	1.6
Viral hepatitis	70	infectious diseases	7.39E-11	1.8
Malignant neoplasm, other	195.1	neoplasms	9.37E-11	1.6
Postinflammatory pulmonary fibrosis	502	respiratory	9.45E-11	2.2
Hx of malignant neoplasm of oral cavity and pharynx	149.5	neoplasms	1E-10	2.0
Chronic pain	338.2	neurological	1.09E-10	1.7
Unstable angina (intermediate coronary syndrome)	411.1	circulatory system	1.34E-10	1.6
Radiotherapy	196	neoplasms	1.35E-10	1.7
Kidney replaced by transpant	587	genitourinary	2.58E-10	0.6
Anxiety disorder	300.1	mental disorders	3.21E-10	1.3
Heart failure with reduced EF [Systolic or combined heart failure]	428.3	circulatory system	4.33E-10	1.5
Acute posthemorrhagic anemia	285.1	hematopoietic	8.3E-10	1.4
Electrolyte imbalance	276.1	endocrine/metabolic	1.33E-09	1.2
Major depressive disorder	296.22	mental disorders	1.36E-09	1.5
Posttraumatic stress disorder	300.9	mental disorders	1.37E-09	2.3
Hypotension NOS	458.9	circulatory system	2.45E-09	1.4
Hypovolemia	276.5	endocrine/metabolic	2.86E-09	1.3
Hypotension	458	circulatory system	2.93E-09	1.3
Suicidal ideation or attempt	297	mental disorders	3.12E-09	2.6
Atrial fibrillation	427.21	circulatory system	3.28E-09	1.3
Secondary malignant neoplasm of liver	198.4	neoplasms	4.61E-09	1.5
Pulmonary congestion and hypostasis	503	respiratory	6.56E-09	1.5
Cancer of prostate	185	neoplasms	9.65E-09	0.7
Suicidal ideation	297.1	mental disorders	1.84E-08	3.0
Chronic pulmonary heart disease	415.2	circulatory system	2.02E-08	1.5
Erectile dysfunction [ED]	605	genitourinary	2.11E-08	0.7
Atrial fibrillation and flutter	427.2	circulatory system	2.71E-08	1.3
Heart failure NOS	428.2	circulatory system	2.77E-08	1.5
Pulmonary heart disease	415	circulatory system	3.19E-08	1.4
Alteration of consciousness	291.8	mental disorders	5.86E-08	1.4
Solitary pulmonary nodule	514.2	respiratory	7.4E-08	2.0
Effects radiation NOS	990	injuries & poisonings	8.88E-08	1.7
Other specified peripheral vascular diseases	443.8	circulatory system	9.9E-08	3.1
Angina pectoris	411.3	circulatory system	1.63E-07	1.5
Shock	797	symptoms	2.7E-07	1.5

Description	Phecode	group	P-value	OR
Schizophrenia	295.1	mental disorders	3.24E-07	2.9
Sepsis and SIRS	994	injuries & poisonings	5.08E-07	1.3
Personality disorders	301	mental disorders	5.32E-07	2.8
Other forms of chronic heart disease	414	circulatory system	7.05E-07	1.3
Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	457	circulatory system	7.36E-07	1.4
Other specified nonpsychotic and/or transient mental disorders	291	mental disorders	7.9E-07	1.4
Arterial embolism and thrombosis	444	circulatory system	8.33E-07	1.8
Paroxysmal ventricular tachycardia	427.12	circulatory system	9.01E-07	1.4
Peptic ulcer, site unspecified	531.4	digestive	9.11E-07	2.1
Human immunodeficiency virus [HIV] disease	71	infectious diseases	1.05E-06	2.2
Cardiac dysrhythmias	427	circulatory system	1.95E-06	1.2
Dysphagia	532	digestive	1.95E-06	1.3
Antisocial/borderline personality disorder	301.2	mental disorders	1.97E-06	3.3
Hearing loss	389	sense organs	2.1E-06	0.8
Myopia	367.1	sense organs	2.31E-06	0.6
Hypopotassemia	276.14	endocrine/metaboli c	2.5E-06	1.2
Chronic obstructive asthma with exacerbation	495.11	respiratory	2.53E-06	3.6
HIV infection, symptomatic	71.1	infectious diseases	3.09E-06	2.2
Cardiac defibrillator in situ	426.92	circulatory system	3.33E-06	1.5
Atherosclerosis of aorta	440.9	circulatory system	3.66E-06	1.8
Other pulmonary inflamation or edema	505	respiratory	4.41E-06	1.7
Respiratory insufficiency	509.2	respiratory	4.46E-06	1.5
End stage renal disease	585.32	genitourinary	5.34E-06	0.7
Anorexia	260.6	endocrine/metaboli c	5.61E-06	1.6
Asthma	495	respiratory	6.64E-06	1.3
Benign neoplasm of skin	216	neoplasms	7.26E-06	0.7
Cardiomyopathy	425	circulatory system	7.73E-06	1.3
Hyperplasia of prostate	600	genitourinary	9.43E-06	0.7
Nausea and vomiting	789	symptoms	9.55E-06	1.2
Heart valve replaced	395.6	circulatory system	1.04E-05	1.5
Anemia in neoplastic disease	285.22	hematopoietic	1.07E-05	1.4
Transient cerebral ischemia	433.31	circulatory system	1.08E-05	1.3
Atherosclerosis of native arteries of the extremities with ulceration or gangrene	440.21	circulatory system	1.09E-05	1.9
Sepsis	994.2	injuries & poisonings	1.09E-05	1.3
Cerebral ischemia	433.3	circulatory system	1.11E-05	1.3

Description	**Phecode**	**group**	**P-value**	**OR**
Encounter for long-term (current) use of aspirin	457.3	circulatory system	1.14E-05	1.4
Cardiogenic shock	797.1	symptoms	1.17E-05	1.9
Primary/intrinsic cardiomyopathies	425.1	circulatory system	1.18E-05	1.3
Abnormal sputum	516	respiratory	1.27E-05	1.5
Thrombocytopenia	287.3	hematopoietic	1.69E-05	1.3
Other alveolar and parietoalveolar pneumonopathy	504	respiratory	1.8E-05	2.0
Cardiomegaly	416	circulatory system	1.9E-05	1.2
Cancer of stomach	151	neoplasms	2E-05	1.7
Schizophrenia and other psychotic disorders	295	mental disorders	2.04E-05	1.6
Other anemias	285	hematopoietic	2.52E-05	1.1
Cardiac pacemaker/device in situ	426.9	circulatory system	2.82E-05	1.3
Purpura and other hemorrhagic conditions	287	hematopoietic	2.86E-05	1.3
Altered mental status	292.4	mental disorders	2.88E-05	1.3
Atherosclerosis of renal artery	440.1	circulatory system	3.05E-05	1.8
Spondylosis and allied disorders	721	musculoskeletal	3.14E-05	1.3
Benign neoplasm of brain, cranial nerves, meninges	225.1	neoplasms	3.2E-05	0.6

APPENDIX D

Significant results with pack-year tobacco exposure via SHAPES

Description	**Phecode**	**Group**	**p-value**
Chronic airway obstruction	496	Respiratory	0*
Tobacco use disorder	318	mental disorders	0*
Chronic bronchitis	496.2	Respiratory	1.5E-208
Shortness of breath	512.7	Respiratory	3.2E-204
Obstructive chronic bronchitis	496.21	Respiratory	7.3E-200
Cancer within the respiratory system	165	Neoplasms	7.5E-194
Cancer of bronchus; lung	165.1	Neoplasms	3.8E-192
Emphysema	496.1	Respiratory	2.9E-186
Other dyspnea	512.9	Respiratory	1.3E-175
Pulmonary collapse; interstitial and compensatory emphysema	508	Respiratory	2.4E-174
Other symptoms of respiratory system	512	Respiratory	7.3E-172
Respiratory failure, insufficiency, arrest	509	respiratory	1.4E-166
Other diseases of lung	510	respiratory	6.2E-160
Coronary atherosclerosis	411.4	circulatory system	2.8E-154
Ischemic Heart Disease	411	circulatory system	2.4E-153
Respiratory failure	509.1	respiratory	4.4E-153
Pleurisy; pleural effusion	507	respiratory	1.9E-145
Myocardial infarction	411.2	circulatory system	1.8E-139
Atherosclerosis	440	circulatory system	8.5E-137
Nonspecific chest pain	418	circulatory system	2.2E-136
Pneumonia	480	respiratory	3E-133
Disorders of fluid, electrolyte, and acid-base balance	276	endocrine/metabolic	2.1E-132
Peripheral vascular disease, unspecified	443.9	circulatory system	6.5E-121
Electrolyte imbalance	276.1	endocrine/metabolic	1.6E-119
Peripheral vascular disease	443	circulatory system	1.2E-117
Cough	512.8	respiratory	2.9E-117
Substance addiction and disorders	316	mental disorders	5.4E-110
Hypovolemia	276.5	endocrine/metabolic	4.1E-105
Atherosclerosis of the extremities	440.2	circulatory system	5.2E-103
Other diseases of respiratory system, not elsewhere classified	519	respiratory	7.7E-103
Symptoms involving respiratory system and other chest symptoms	519.9	respiratory	9.7E-103
Empyema and pneumothorax	506	respiratory	1.8E-102

Description	Phecode	Group	p-value
Alcohol-related disorders	317	mental disorders	7.29E-94
Tachycardia NOS	427.7	circulatory system	1.5E-93
Abnormal findings examination of lungs	514	respiratory	4.48E-91
Mood disorders	296	mental disorders	2.06E-88
Other anemias	285	hematopoietic	1.76E-87
Depression	296.2	mental disorders	8.23E-87
Cardiac dysrhythmias	427	circulatory system	5.13E-86
Hypotension	458	circulatory system	1.34E-85
Secondary malignant neoplasm	198	neoplasms	4.01E-85
Other chronic ischemic heart disease, unspecified	411.8	circulatory system	1.11E-84
Congestive heart failure; nonhypertensive	428	circulatory system	2.17E-83
Acute pain	338.1	neurological	4.13E-82
Hypopotassemia	276.14	endocrine/metabolic	1.67E-81
Unstable angina (intermediate coronary syndrome)	411.1	circulatory system	4.49E-78
Dysphagia	532	digestive	8.24E-77
Acute renal failure	585.1	genitourinary	1.62E-76
Congestive heart failure (CHF) NOS	428.1	circulatory system	3.36E-76
Diseases of white blood cells	288	hematopoietic	2.25E-75
Occlusion and stenosis of precerebral arteries	433.1	circulatory system	3.72E-75
Cerebrovascular disease	433	circulatory system	8.04E-75
Atherosclerosis of native arteries of the extremities with intermittent claudication	440.22	circulatory system	1.37E-74
Hypotension NOS	458.9	circulatory system	4.34E-74
Chronic obstructive asthma	495.1	respiratory	2.2E-73
Anxiety, phobic and dissociative disorders	300	mental disorders	3.6E-72
Asthma	495	respiratory	5.13E-71
Alcoholism	317.1	mental disorders	8.68E-71
Angina pectoris	411.3	circulatory system	7.47E-68
Acid-base balance disorder	276.4	endocrine/metabolic	2.79E-67
Diseases of esophagus	530	digestive	1.41E-65
Other aneurysm	442	circulatory system	6.21E-65
Nausea and vomiting	789	symptoms	6.76E-65
Renal failure	585	genitourinary	1.36E-64
Esophagitis, GERD and related diseases	530.1	digestive	1.31E-62
Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	457	circulatory system	4.17E-62
Cancer, suspected or other	195	neoplasms	8.7E-62
Elevated white blood cell count	288.2	hematopoietic	1.27E-60
Hyperpotassemia	276.13	endocrine/metabolic	2.87E-60
Cancer of larynx, pharynx, nasal cavities	149	neoplasms	3.34E-60
Pulmonary congestion and hypostasis	503	respiratory	4.31E-60
Anxiety disorder	300.1	mental disorders	7.32E-60

Description	**Phecode**	**Group**	**p-value**
Malaise and fatigue	798	symptoms	1.09E-59
Other forms of chronic heart disease	414	circulatory system	1.74E-59
Acute posthemorrhagic anemia	285.1	hematopoietic	2.87E-59
Protein-calorie malnutrition	260	endocrine/metabolic	4.06E-59
Hypertension	401	circulatory system	1.17E-58
Essential hypertension	401.1	circulatory system	1.79E-58
GERD	530.11	digestive	1.85E-58
Aortic aneurysm	442.1	circulatory system	2.87E-58
Pulmonary insufficiency or respiratory failure following trauma and surgery	509.3	respiratory	1.11E-57
Painful respiration	512.2	respiratory	1.95E-57
Atrial fibrillation	427.21	circulatory system	6.43E-57
Cancer of larynx	149.4	neoplasms	6.51E-57
Atrial fibrillation and flutter	427.2	circulatory system	1.46E-56
Fever of unknown origin	783	symptoms	1.95E-56
Chronic pain	338.2	neurological	1.38E-55
Bacterial pneumonia	480.1	respiratory	3.49E-55
Acidosis	276.41	endocrine/metabolic	9.27E-55
Dependence on respirator [Ventilator] or supplemental oxygen	509.8	respiratory	1.92E-54
Abdominal aortic aneurysm	442.11	circulatory system	1.96E-54
Lymphadenitis	289.4	hematopoietic	1.97E-54
Chemotherapy	197	neoplasms	2.38E-54
Encounter for long-term (current) use of aspirin	457.3	circulatory system	1.19E-53
Hypertensive heart and/or renal disease	401.2	circulatory system	3.75E-53
Type 2 diabetes	250.2	endocrine/metabolic	1.29E-52
Hyperlipidemia	272.1	endocrine/metabolic	7.39E-52
Disorders of lipoid metabolism	272	endocrine/metabolic	1.7E-51
Hyposmolality and/or hyponatremia	276.12	endocrine/metabolic	2.04E-51
Wheezing	512.1	respiratory	3.05E-51
Other disorders of the kidney and ureters	586	genitourinary	4.66E-51
Septicemia	38	infectious diseases	1.73E-50
Diabetes mellitus	250	endocrine/metabolic	3.84E-50
Secondary malignancy of lymph nodes	198.1	neoplasms	7.96E-50
Anemia of chronic disease	285.2	hematopoietic	3.15E-49
Bacterial infection NOS	41	infectious diseases	4.97E-49
Hypertensive chronic kidney disease	401.22	circulatory system	1.5E-47
Fluid overload	276.6	endocrine/metabolic	2.02E-47
Cardiomegaly	416	circulatory system	5.62E-47
Altered mental status	292.4	mental disorders	1.65E-45
Pulmonary heart disease	415	circulatory system	2.98E-44
Secondary malignancy of respiratory organs	198.2	neoplasms	3.6E-44

Description	Phecode	Group	p-value
Heart failure with preserved EF [Diastolic heart failure]	428.4	circulatory system	1.07E-43
Heart failure with reduced EF [Systolic or combined heart failure]	428.3	circulatory system	2.3E-43
Edema	782.3	symptoms	2.91E-43
Heart failure NOS	428.2	circulatory system	2.31E-42
Cardiac conduction disorders	426	circulatory system	5.81E-42
Major depressive disorder	296.22	mental disorders	1.02E-41
Sepsis and SIRS	994	injuries & poisonings	1.15E-41
Ill-defined descriptions and complications of heart disease	429	circulatory system	1.49E-41
Other specified nonpsychotic and/or transient mental disorders	291	mental disorders	2.32E-41
Other disorders of circulatory system	459	circulatory system	3.8E-41
Arrhythmia (cardiac) NOS	427.5	circulatory system	1.74E-40
Alteration of consciousness	291.8	mental disorders	2.29E-40
Malignant neoplasm, other	195.1	neoplasms	2.78E-40
Cancer of mouth	145	neoplasms	1.13E-39
Abnormal sputum	516	respiratory	1.14E-39
Arterial embolism and thrombosis	444	circulatory system	1.52E-39
Chronic renal failure [CKD]	585.3	genitourinary	1.95E-39
Sepsis	994.2	injuries & poisonings	4.03E-39
Mixed hyperlipidemia	272.13	endocrine/metabolic	1.91E-38
Chronic pulmonary heart disease	415.2	circulatory system	6.66E-38
Hypertensive heart disease	401.21	circulatory system	1E-37
Abdominal pain	785	symptoms	2.57E-37
Hemoptysis	516.1	respiratory	8.71E-37
Paroxysmal ventricular tachycardia	427.12	circulatory system	9.52E-37
Other specified cardiac dysrhythmias	427.3	circulatory system	1.05E-36
Paroxysmal tachycardia, unspecified	427.1	circulatory system	1.87E-36
Other pulmonary inflamation or edema	505	respiratory	2.06E-36
Type 1 diabetes	250.1	endocrine/metabolic	2.82E-36
Circulatory disease NEC	459.9	circulatory system	5.09E-36
Neurological disorders	292	mental disorders	5.4E-36
Bipolar	296.1	mental disorders	1.3E-35
Type 2 diabetes with neurological manifestations	250.24	endocrine/metabolic	1.35E-35
Staphylococcus infections	41.1	infectious diseases	1.42E-35
Insulin pump user	250.3	endocrine/metabolic	1.95E-35
Iron deficiency anemias	280	hematopoietic	6.82E-35
Cerebral ischemia	433.3	circulatory system	1.33E-34
Respiratory insufficiency	509.2	respiratory	1.7E-34

Description	Phecode	Group	p-value
Back pain	760	symptoms	3.16E-34
Solitary pulmonary nodule	514.2	respiratory	3.95E-34
Transient cerebral ischemia	433.31	circulatory system	1.76E-33
Bronchitis	497	respiratory	2.23E-33
Atherosclerosis of aorta	440.9	circulatory system	3.19E-33
Superficial cellulitis and abscess	681	dermatologic	1.01E-32
Chronic ulcer of skin	707	dermatologic	1.22E-32
Hypercholesterolemia	272.11	endocrine/metabolic	2.47E-32
Other abnormal glucose	250.42	endocrine/metabolic	3.22E-32
Secondary malignancy of brain/spine	198.5	neoplasms	8.43E-32
Anorexia	260.6	endocrine/metabolic	1.17E-31
Hx of malignant neoplasm of oral cavity and pharynx	149.5	neoplasms	2.23E-31
Cardiac pacemaker/device in situ	426.9	circulatory system	6.81E-31
Complications of cardiac/vascular device, implant, and graft	854	injuries & poisonings	7.48E-31
Iron deficiency anemias, unspecified or not due to blood loss	280.1	hematopoietic	1.88E-30
Other venous embolism and thrombosis	452	circulatory system	6.12E-30
Abnormal glucose	250.4	endocrine/metabolic	1.28E-29
Cancer of oropharynx	149.1	neoplasms	2.17E-29
Gastrointestinal hemorrhage	578	digestive	3.18E-29
Postinflammatory pulmonary fibrosis	502	respiratory	3.25E-29
Pneumonitis due to inhalation of food or vomitus	501	respiratory	3.79E-29
Type 2 diabetes with renal manifestations	250.22	endocrine/metabolic	4.48E-29
Polyneuropathy in diabetes	250.6	endocrine/metabolic	1.1E-28
Dysthymic disorder	300.4	mental disorders	3.21E-28
Acute bronchitis and bronchiolitis	483	respiratory	4.3E-28
Bacteremia	38.3	infectious diseases	1.29E-27
Swelling, mass, or lump in head and neck [Space-occupying lesion, intracranial NOS]	293.1	mental disorders	1.56E-27
Cardiac defibrillator in situ	426.92	circulatory system	1.62E-27
Respiratory abnormalities	513	respiratory	6.36E-27
Chronic obstructive asthma with exacerbation	495.11	respiratory	9.4E-27
Swelling of limb	771.1	symptoms	1.68E-26
Atherosclerosis of renal artery	440.1	circulatory system	2.29E-26
Coagulation defects	286	hematopoietic	2.5E-26
Posttraumatic stress disorder	300.9	mental disorders	2.75E-26
Other specified peripheral vascular diseases	443.8	circulatory system	8.93E-26
ASCVD	414.2	circulatory system	9.65E-26
Spondylosis and allied disorders	721	musculoskeletal	1.52E-25
Chronic Kidney Disease, Stage III	585.33	genitourinary	1.6E-25
Cervicalgia	761	symptoms	1.72E-25

Description	Phecode	Group	p-value
Atherosclerosis of native arteries of the extremities with ulceration or gangrene	440.21	circulatory system	1.84E-25
Encounter for long-term (current) use of anticoagulants	286.2	hematopoietic	2.13E-25
Renal failure NOS	585.2	genitourinary	3.24E-25
Spondylosis without myelopathy	721.1	musculoskeletal	1.14E-24
Heart valve disorders	395	circulatory system	1.76E-24
Syncope and collapse	788	symptoms	2.54E-24
Traumatic and surgical subcutaneous emphysema	958.2	injuries & poisonings	4.51E-24
Other peripheral nerve disorders	351	neurological	4.92E-24
Secondary malignancy of bone	198.6	neoplasms	5.95E-24
Septic shock	994.21	injuries & poisonings	6.58E-24
Neoplasm of uncertain behavior	199	neoplasms	9.49E-24
Occlusion of cerebral arteries	433.2	circulatory system	1.33E-23
Palpitations	427.9	circulatory system	4.02E-23
Anemia in neoplastic disease	285.22	hematopoietic	4.23E-23
Candidiasis	112	infectious diseases	5.77E-23
Intervertebral disc disorders	722	musculoskeletal	7.85E-23
Diverticulosis	562.1	digestive	1.01E-22
Cancer of tongue	145.2	neoplasms	1.08E-22
Diseases of the larynx and vocal cords	473	respiratory	1.14E-22
Infection with drug-resistant microorganisms	41.9	infectious diseases	1.15E-22
Hemorrhage of gastrointestinal tract	578.9	digestive	2.63E-22
severe protein-calorie malnutrition	260.2	endocrine/metabolic	3.41E-22
Cerebral artery occlusion, with cerebral infarction	433.21	circulatory system	5.89E-22
Chronic ulcer of leg or foot	707.2	dermatologic	6.01E-22
Blood in stool	578.2	digestive	6.13E-22
Abnormal serum enzyme levels	573.9	digestive	7.17E-22
Cardiomyopathy	425	circulatory system	8.05E-22
Urinary tract infection	591	genitourinary	1.21E-21
Cardiac pacemaker in situ	426.91	circulatory system	3.06E-21
Acute, but ill-defined cerebrovascular disease	433.6	circulatory system	3.55E-21
Adverse drug events and drug allergies	979	injuries & poisonings	4.15E-21
Stricture and stenosis of esophagus	530.3	digestive	6.02E-21
Diverticulosis and diverticulitis	562	digestive	6.45E-21
Shock	797	symptoms	7.03E-21
Atrial flutter	427.22	circulatory system	1.22E-20
Decubitus ulcer	707.1	dermatologic	1.5E-20
Chronic Kidney Disease, Stage IV	585.34	genitourinary	2.11E-20
Symptoms involving cardiovascular system	429.3	circulatory system	2.5E-20

Description	Phecode	Group	p-value
Alcoholic liver damage	317.11	mental disorders	4.18E-20
Primary/intrinsic cardiomyopathies	425.1	circulatory system	4.54E-20
Benign neoplasm of colon	208	neoplasms	4.56E-20
Methicillin sensitive Staphylococcus aureus	41.11	infectious diseases	4.64E-20
Encounter for long-term (current) use of antibiotics	980	infectious diseases	6.49E-20
Anemia in chronic kidney disease	285.21	hematopoietic	7.44E-20
Cellulitis and abscess of leg, except foot	681.5	dermatologic	8.51E-20
Other hypertensive complications	401.3	circulatory system	9.03E-20
Effects radiation NOS	990	injuries & poisonings	1.19E-19
Precordial pain	418.1	circulatory system	1.45E-19
Bundle branch block	426.3	circulatory system	1.79E-19
Generalized anxiety disorder	300.11	mental disorders	3.69E-19
Degeneration of intervertebral disc	722.6	musculoskeletal	3.81E-19
Purpura and other hemorrhagic conditions	287	hematopoietic	4.77E-19
Type 1 diabetes with neurological manifestations	250.14	endocrine/metabolic	7.79E-19
End stage renal disease	585.32	genitourinary	9.09E-19
Radiotherapy	196	neoplasms	1.33E-18
Other symptoms/disorders or the urinary system	599	genitourinary	1.52E-18
Certain early complications of trauma or procedure	958	injuries & poisonings	1.57E-18
Sleep apnea	327.3	neurological	1.72E-18
Deep vein thrombosis [DVT]	452.2	circulatory system	1.75E-18
Obstructive sleep apnea	327.32	neurological	1.9E-18
Bronchopneumonia and lung abscess	480.5	respiratory	1.95E-18
Voice disturbance	473.4	respiratory	4.36E-18
Diabetes type 2 with peripheral circulatory disorders	250.25	endocrine/metabolic	6.29E-18
Thrombocytopenia	287.3	hematopoietic	1.47E-17
Postoperative infection	80	infectious diseases	1.64E-17
Agorophobia, social phobia, and panic disorder	300.12	mental disorders	1.95E-17
Nonrheumatic mitral valve disorders	395.1	circulatory system	2.43E-17
Abnormal coagulation profile	286.9	hematopoietic	3.69E-17
Suicidal ideation or attempt	297	mental disorders	4.53E-17
Disorders of mineral metabolism	275	endocrine/metabolic	4.53E-17
Diseases of the oral soft tissues, excluding lesions specific for gingiva and tongue	528	digestive	4.54E-17
Cerebral atherosclerosis	433.12	circulatory system	5.23E-17
Adult failure to thrive	260.3	endocrine/metabolic	9.98E-17
Aplastic anemia	284	hematopoietic	1.11E-16

Description	Phecode	Group	p-value
Renal dialysis	585.31	genitourinary	1.14E-16
Streptococcus infection	41.2	infectious diseases	3.14E-16
Schizophrenia and other psychotic disorders	295	mental disorders	5.49E-16
Left bundle branch block	426.32	circulatory system	5.87E-16
Hereditary and idiopathic peripheral neuropathy	356	neurological	6.58E-16
Sleep disorders	327	neurological	6.74E-16
Cardiac arrest and ventricular fibrillation	427.4	circulatory system	7.8E-16
Insomnia	327.4	neurological	1.35E-15
Neutropenia	288.11	hematopoietic	1.6E-15
Other and unspecified coagulation defects	286.7	hematopoietic	1.66E-15
Orthostatic hypotension	458.1	circulatory system	2.04E-15
Disturbance of skin sensation	687.4	dermatologic	2.06E-15
Abnormal electrocardiogram [ECG] [EKG]	426.7	circulatory system	2.19E-15
Cholelithiasis and cholecystitis	574	digestive	2.21E-15
Peptic ulcer (excl. esophageal)	531	digestive	2.33E-15
Convulsions	345.3	neurological	2.55E-15
Overweight, obesity and other hyperalimentation	278	endocrine/metabolic	2.59E-15
Decreased white blood cell count	288.1	hematopoietic	3.44E-15
Other disorders of stomach and duodenum	537	digestive	3.61E-15
Viral hepatitis	70	infectious diseases	3.84E-15
Carditis	420	circulatory system	4.35E-15
Cellulitis and abscess of trunk	681.7	dermatologic	6.05E-15
Other disorders of arteries and arterioles	447	circulatory system	8.64E-15
Encounter for long-term (current) use of antiplatelets/antithrombotics	457.2	circulatory system	9.14E-15
Cachexia	260.1	endocrine/metabolic	1.13E-14
E. coli	41.4	infectious diseases	1.25E-14
Atrioventricular [AV] block	426.2	circulatory system	1.35E-14
Sinoatrial node dysfunction (Bradycardia)	427.8	circulatory system	1.37E-14
Methicillin resistant Staphylococcus aureus	41.12	infectious diseases	1.56E-14
Embolism and thrombosis of abdominal aorta	444.2	circulatory system	1.69E-14
Pancytopenia	284.1	hematopoietic	1.99E-14
Nephritis; nephrosis; renal sclerosis	580	genitourinary	2.36E-14
Mycoses	117	infectious diseases	2.48E-14
Systemic inflammatory response syndrome (SIRS)	994.1	injuries & poisonings	2.52E-14
Viral hepatitis C	70.3	infectious diseases	3.35E-14
Hematuria	593	genitourinary	3.63E-14
Epilepsy, recurrent seizures, convulsions	345	neurological	5.71E-14
Asthma with exacerbation	495.2	respiratory	6.18E-14
Adjustment reaction	304	mental disorders	6.38E-14
Alkalosis	276.42	endocrine/metabolic	9E-14

Description	Phecode	Group	p-value
Personality disorders	301	mental disorders	9.43E-14
Peptic ulcer, site unspecified	531.4	digestive	1.04E-13
Obesity	278.1	endocrine/metabolic	1.13E-13
Type 2 diabetes with ophthalmic manifestations	250.23	endocrine/metabolic	1.17E-13
Other conditions of brain, NOS	348.9	neurological	1.31E-13
Late effects of cerebrovascular disease	433.8	circulatory system	1.34E-13
Noninfectious gastroenteritis	558	digestive	1.67E-13
Disorders of magnesium metabolism	275.3	endocrine/metabolic	1.95E-13
Infection/inflammation of internal prosthetic device; implant; and graft	81	infectious diseases	2.15E-13
Gangrene	791	symptoms	2.58E-13
Intestinal infection	8	infectious diseases	2.72E-13
Other symptoms involving abdomen and pelvis	579	digestive	2.74E-13
Acute pulmonary heart disease	415.1	circulatory system	3.3E-13
Iron deficiency anemia secondary to blood loss (chronic)	280.2	hematopoietic	4.4E-13
Cancer of the mouth floor	145.5	neoplasms	5E-13
Disorders of function of stomach	536	digestive	6.38E-13
Nonrheumatic aortic valve disorders	395.2	circulatory system	6.54E-13
Diaphragmatic hernia	550.2	digestive	6.56E-13
Hyperventilation	513.4	respiratory	6.79E-13
Suicidal ideation	297.1	mental disorders	7.42E-13
Poisoning by antibiotics	960	injuries & poisonings	8.18E-13
Phlebitis and thrombophlebitis	451	circulatory system	9.19E-13
Secondary malignant neoplasm of liver	198.4	neoplasms	9.4E-13
Other infectious and parasitic diseases	136	infectious diseases	1.03E-12
Disorders of sweat glands	705	dermatologic	1.2E-12
Other alveolar and parietoalveolar pneumonopathy	504	respiratory	1.37E-12
Other local infections of skin and subcutaneous tissue	686	dermatologic	1.72E-12
Occlusion of cerebral arteries, with cerebral infarction	433.11	circulatory system	1.82E-12
Pericarditis	420.2	circulatory system	1.83E-12
Human immunodeficiency virus [HIV] disease	71	infectious diseases	2.11E-12
Abdominal hernia	550	digestive	2.19E-12
Hypoglycemia	251.1	endocrine/metabolic	2.44E-12
Other disorders of intestine	569	digestive	2.62E-12
Disorders resulting from impaired renal function	588	genitourinary	2.71E-12
Stricture of artery	447.1	circulatory system	2.71E-12

Description	Phecode	Group	p-value
Cerebral aneurysm	433.5	circulatory system	2.79E-12
Osteoarthrosis NOS	740.9	musculoskeletal	2.82E-12
Somatoform disorder	303.4	mental disorders	3.22E-12
Complications of gastrostomy, colostomy and enterostomy	536.7	digestive	4.11E-12
Dermatophytosis of nail	110.11	infectious diseases	4.12E-12
Other conditions of brain	348	neurological	5.23E-12
Pneumococcal pneumonia	480.11	respiratory	5.8E-12
Incisional hernia	550.6	digestive	6.04E-12
Other disorders of cervical region	723	musculoskeletal	6.26E-12
Hemorrhage or hematoma complicating a procedure	850	injuries & poisonings	6.45E-12
Diseases of hard tissues of teeth	521	digestive	6.51E-12
Cholelithiasis	574.1	digestive	6.88E-12
Hyperosmolality and/or hypernatremia	276.11	endocrine/metabolic	7.89E-12
Influenza	481	respiratory	9.94E-12
Abnormal function study of cardiovascular system	429.2	circulatory system	1.06E-11
Stomatitis and mucositis	528.1	digestive	1.07E-11
Allergy/adverse effect of penicillin	960.2	injuries & poisonings	1.22E-11
Bronchiectasis	496.3	respiratory	1.28E-11
Delirium due to conditions classified elsewhere	290.2	mental disorders	1.3E-11
Gram positive septicemia	38.2	infectious diseases	1.34E-11
HIV infection, symptomatic	71.1	infectious diseases	1.37E-11
Pyelonephritis	590	genitourinary	1.38E-11
Type 1 diabetes with renal manifestations	250.12	endocrine/metabolic	1.43E-11
Cellulitis and abscess of arm/hand	681.3	dermatologic	1.53E-11
Chronic liver disease and cirrhosis	571	digestive	1.57E-11
Cardiogenic shock	797.1	symptoms	1.81E-11
Pseudomonal pneumonia	480.12	respiratory	1.89E-11
Idiopathic fibrosing alveolitis	504.1	respiratory	2.09E-11
Contusion	916	injuries & poisonings	2.23E-11
Arterial embolism and thrombosis of lower extremity artery	444.1	circulatory system	2.31E-11
Paroxysmal supraventricular tachycardia	427.11	circulatory system	2.97E-11
Pulmonary embolism and infarction, acute	415.11	circulatory system	3.28E-11
Cancer of esophagus	150	neoplasms	3.68E-11
Chronic pain syndrome	355.1	neurological	3.92E-11
Osteoporosis, osteopenia and pathological fracture	743	musculoskeletal	4.97E-11
Dental caries	521.1	digestive	5.3E-11

Description	Phecode	Group	p-value
Poisoning by hormones and synthetic substitutes	962	injuries & poisonings	5.44E-11
Heart valve replaced	395.6	circulatory system	5.57E-11
Dermatophytosis / Dermatomycosis	110	infectious diseases	6.54E-11
Myopathy	359.2	neurological	8.02E-11
Ascites (non malignant)	572	digestive	9.14E-11
Heart transplant/surgery	429.1	circulatory system	9.89E-11
Rheumatic disease of the heart valves	394	circulatory system	1.01E-10
Other upper respiratory disease	479	respiratory	1.05E-10
Diseases of pancreas	577	digestive	1.25E-10
MRSA pneumonia	480.13	respiratory	1.33E-10
Aphasia/speech disturbance	292.1	mental disorders	1.74E-10
Nerve root and plexus disorders	353	neurological	1.88E-10
Heartburn	530.9	digestive	1.92E-10
Senile cataract	366.2	sense organs	1.99E-10
Intestinal obstruction without mention of hernia	560	digestive	2.01E-10
Other chronic nonalcoholic liver disease	571.5	digestive	2.07E-10
Retention of urine	599.2	genitourinary	3.42E-10
Frequency of urination and polyuria	599.5	genitourinary	3.51E-10
Osteoporosis NOS	743.11	musculoskeletal	3.57E-10
Muscular dystrophies and other myopathies	359	neurological	3.65E-10
Diabetic retinopathy	250.7	endocrine/metabolic	3.7E-10
Spasm of muscle	772.2	symptoms	3.8E-10
Hemorrhage of rectum and anus	578.8	digestive	4.22E-10
Atrioventricular block, complete	426.24	circulatory system	4.43E-10
Pain in joint	745	musculoskeletal	4.57E-10
Morbid obesity	278.11	endocrine/metabolic	4.81E-10
Cellulitis and abscess of foot, toe	681.6	dermatologic	4.83E-10
Cataract	366	sense organs	5.02E-10
Iatrogenic hypotension	458.2	circulatory system	5.39E-10
Delirium dementia and amnestic and other cognitive disorders	290	mental disorders	5.46E-10
Cyst of kidney, acquired	586.2	genitourinary	5.6E-10
Fracture of lower limb	800	injuries & poisonings	5.78E-10
Cancer of prostate	185	neoplasms	5.91E-10
Psychosis	295.3	mental disorders	6E-10
Osteomyelitis, periostitis, and other infections involving bone	710	musculoskeletal	7.83E-10
Vascular insufficiency of intestine	441	circulatory system	8E-10
Other disorders of pancreatic internal secretion	251	endocrine/metabolic	8.39E-10
Acute osteomyelitis	710.11	musculoskeletal	9.84E-10
Antisocial/borderline personality disorder	301.2	mental disorders	1.2E-09

Description	Phecode	Group	p-value
Intestinal infection due to C. difficile	8.52	infectious diseases	1.23E-09
Encephalopathy, not elsewhere classified	348.8	neurological	1.24E-09
Malignant neoplasm of bladder	189.21	neoplasms	1.25E-09
Complications of transplants and reattached limbs	851	injuries & poisonings	1.28E-09
Vitamin B-complex deficiencies	261.2	endocrine/metabolic	1.32E-09
Acute pancreatitis	577.1	digestive	1.49E-09
Ventral hernia	550.5	digestive	1.51E-09
Dysuria	599.3	genitourinary	1.59E-09
Open wounds of extremities	871	injuries & poisonings	1.6E-09
Bacterial enteritis	8.5	infectious diseases	1.7E-09
Dermatophytosis	110.1	infectious diseases	1.75E-09
Renal sclerosis, NOS	580.4	genitourinary	1.76E-09
Ventricular fibrillation and flutter	427.41	circulatory system	2.08E-09
Cancer of bladder	189.2	neoplasms	2.08E-09
Chronic venous insufficiency [CVI]	456	circulatory system	2.13E-09
Schizophrenia	295.1	mental disorders	2.15E-09
Spinal stenosis	720	musculoskeletal	2.75E-09
Osteomyelitis	710.1	musculoskeletal	3.31E-09
Psychogenic and somatoform disorders	303	mental disorders	3.68E-09
Tracheostomy complications	519.1	respiratory	4.84E-09
Primary pulmonary hypertension	415.21	circulatory system	5.05E-09
Cancer of hypopharynx	149.3	neoplasms	5.27E-09
Disturbance of salivary secretion	527.7	digestive	5.29E-09
Suicide or self-inflicted injury	297.2	mental disorders	5.76E-09
Anal and rectal conditions	565	digestive	6.32E-09
Nephritis and nephropathy in diseases classified elsewhere	580.31	genitourinary	6.36E-09
Chronic sinusitis	475	respiratory	6.53E-09
Cellulitis and abscess of face/neck	681.2	dermatologic	6.89E-09
Other disorders of liver	573	digestive	8.85E-09
Other arthropathies	716	musculoskeletal	9.02E-09
Nonspecific elevation of levels of transaminase or lactic acid dehydrogenase [LDH]	573.6	digestive	1.04E-08
Disorders of phosphorus metabolism	275.53	endocrine/metabolic	1.14E-08
Opiates and related narcotics causing adverse effects in therapeutic use	965.1	injuries & poisonings	1.24E-08
Secondary/extrinsic cardiomyopathies	425.2	circulatory system	1.29E-08
Cerebral degeneration, unspecified	331.9	neurological	1.81E-08
Adrenal cortical steroids causing adverse effects in therapeutic use	962.1	injuries & poisonings	2E-08
Hidradenitis	705.3	dermatologic	2.04E-08

Description	Phecode	Group	p-value
Lung transplant	510.2	respiratory	2.11E-08
Right bundle branch block	426.31	circulatory system	2.16E-08
Arthropathy NOS	716.9	musculoskeletal	2.18E-08
Osteoporosis	743.1	musculoskeletal	2.76E-08
Antineoplastic and immunosuppressive drugs causing adverse effects	963.1	injuries & poisonings	3.12E-08
Muscle weakness	772.3	symptoms	3.36E-08
Other deficiency anemia	281	hematopoietic	3.49E-08
Pneumonia due to fungus (mycoses)	480.3	respiratory	3.83E-08
Premature beats	427.6	circulatory system	4.16E-08
Throat pain	478	respiratory	4.24E-08
Cirrhosis of liver without mention of alcohol	571.51	digestive	4.43E-08
Peritonitis and retroperitoneal infections	567	digestive	4.48E-08
Chronic osteomyelitis	710.12	musculoskeletal	5E-08
Renal osteodystrophy	588.1	genitourinary	5.03E-08
Disorders of calcium/phosphorus metabolism	275.5	endocrine/metabolic	5.14E-08
Nonrheumatic tricuspid valve disorders	395.3	circulatory system	5.61E-08
Disorders of adrenal glands	255	endocrine/metabolic	5.62E-08
Poisoning by primarily systemic agents	963	injuries & poisonings	5.84E-08
Aneurysm and dissection of heart	411.41	circulatory system	6.52E-08
First degree AV block	426.21	circulatory system	6.81E-08
Phlebitis and thrombophlebitis of lower extremities	451.2	circulatory system	6.82E-08
Gastritis and duodenitis	535	digestive	6.84E-08
Unspecified osteomyelitis	710.19	musculoskeletal	6.88E-08
Sprains and strains of back and neck	841	injuries & poisonings	7.8E-08
Benign neoplasm of other parts of digestive system	211	neoplasms	7.96E-08
Calculus of bile duct	574.2	digestive	8.1E-08
Other intestinal obstruction	560.4	digestive	8.66E-08
Cardiac arrest	427.42	circulatory system	8.84E-08
Thoracic or lumbosacral neuritis or radiculitis, unspecified	763	symptoms	9.26E-08
Chronic kidney disease, Stage I or II	585.4	genitourinary	9.94E-08
Corns and callosities	700	dermatologic	1.06E-07
Chronic pericarditis	420.22	circulatory system	1.29E-07
Dizziness and giddiness (Light-headedness and vertigo)	386.9	sense organs	1.32E-07
Neuralgia, neuritis, and radiculitis NOS	766	symptoms	1.38E-07
Fracture of neck of femur	800.1	injuries & poisonings	1.49E-07
Type 1 diabetes with ketoacidosis	250.11	endocrine/metabolic	1.53E-07

Description	Phecode	Group	p-value
Fracture of vertebral column without mention of spinal cord injury	805	injuries & poisonings	1.92E-07
Other diseases of the teeth and supporting structures	525	digestive	1.96E-07
Open wounds of head; neck; and trunk	870	injuries & poisonings	1.97E-07
Personal history of allergy to medicinal agents	977	injuries & poisonings	2.23E-07
Allergies, other	949	injuries & poisonings	2.3E-07
Nephritis and nephropathy without mention of glomerulonephritis	580.3	genitourinary	2.39E-07
Tuberculosis	10	infectious diseases	2.5E-07
Hepatitis NOS	70.9	infectious diseases	2.6E-07
Hemiplegia	342	neurological	2.67E-07
Gastroparesis	536.3	digestive	2.87E-07
Musculoskeletal symptoms referable to limbs	771	symptoms	2.97E-07
Aspergillosis	117.4	infectious diseases	3.02E-07
Type 1 diabetes with ophthalmic manifestations	250.13	endocrine/metabolic	3.27E-07
Rhabdomyolysis	772.4	symptoms	3.61E-07
Symptoms and disorders of the joints	741	musculoskeletal	3.66E-07
Perinatal disorders of digestive system	656.6	pregnancy complications	3.96E-07
Dyshidrosis	705.1	dermatologic	4.06E-07
Other cerebral degenerations	331	neurological	4.27E-07
Inflammatory and toxic neuropathy	357	neurological	4.57E-07
Other diseases of blood and blood-forming organs	289	hematopoietic	4.68E-07
Polycythemia vera, secondary	289.8	hematopoietic	4.89E-07
Cancer of nasopharynx	149.2	neoplasms	5.09E-07
Aneurysm of artery of lower extremity	442.3	circulatory system	5.21E-07
Cancer of urinary organs (incl. kidney and bladder)	189	neoplasms	5.77E-07
Poisoning by psychotropic agents	969	injuries & poisonings	5.77E-07
Fracture of ribs	807	injuries & poisonings	5.82E-07
Disease of tricuspid valve	394.7	circulatory system	5.86E-07
Cardiac complications, not elsewhere classified	429.9	circulatory system	6.3E-07
Vitamin deficiency	261	endocrine/metabolic	8.44E-07
Arthropathy associated with neurological disorders	713.5	musculoskeletal	9.22E-07

Description	Phecode	Group	p-value
Displacement of intervertebral disc	722.1	musculoskeletal	9.44E-07
Fracture of ankle and foot	801	injuries & poisonings	9.61E-07
Anticoagulants causing adverse effects	964.1	injuries & poisonings	1.04E-06
Histoplasmosis	117.1	infectious diseases	1.25E-06
Complication of internal orthopedic device	858	injuries & poisonings	1.26E-06
Gout and other crystal arthropathies	274	endocrine/metabolic	1.42E-06
Other abnormal blood chemistry	790.6	symptoms	1.43E-06
Peripheral enthesopathies and allied syndromes	726	musculoskeletal	1.69E-06
Nerve root lesions	353.2	neurological	1.88E-06
Diseases of the salivary glands	527	digestive	1.88E-06
Gout	274.1	endocrine/metabolic	1.9E-06
Osteopenia or other disorder of bone and cartilage	743.9	musculoskeletal	1.91E-06
Poisoning by anticonvulsants and anti-Parkinsonism drugs	966	injuries & poisonings	1.99E-06
Other disorders of synovium, tendon, and bursa	727	musculoskeletal	2.18E-06
Cancer of of nasal cavities	149.9	neoplasms	2.45E-06
Nutritional marasmus	260.22	endocrine/metabolic	2.45E-06
Acute vascular insufficiency of intestine	441.1	circulatory system	2.46E-06
Kidney replaced by transpant	587	genitourinary	2.48E-06
Other disorders of metabolism	277	endocrine/metabolic	2.49E-06
Abnormality of gait	350.2	neurological	2.54E-06
Deficiency anemias	281.9	hematopoietic	2.55E-06
Colorectal cancer	153	neoplasms	2.64E-06
Cystitis	592.1	genitourinary	2.75E-06
Hemorrhoids	455	circulatory system	2.79E-06
Immunity deficiency	279.1	endocrine/metabolic	2.8E-06
Hemorrhagic disorder due to intrinsic circulating anticoagulants	286.5	hematopoietic	2.99E-06
Paralytic ileus	560.1	digestive	3.31E-06
Cystitis and urethritis	592	genitourinary	3.55E-06
Gram negative septicemia	38.1	infectious diseases	3.59E-06
Diverticulitis	562.2	digestive	3.65E-06
Myalgia and myositis unspecified	770	symptoms	3.78E-06
Other disorders of thyroid	246	endocrine/metabolic	3.88E-06
Benign neoplasm of lip, oral cavity, and pharynx	210	neoplasms	4.17E-06
Poisoning/allergy of sulfonamides	961.1	injuries & poisonings	4.43E-06

Description	Phecode	Group	p-value
Hypoventilation	513.3	respiratory	4.49E-06
Disorders involving the immune mechanism	279	endocrine/metabolic	5.09E-06
Chronic pancreatitis	577.2	digestive	5.64E-06
Other diseases of respiratory system, NEC	519.8	respiratory	5.72E-06
Intracranial hemorrhage	430	circulatory system	5.78E-06
Polyneuropathy due to drugs	316.1	mental disorders	6.14E-06
Viral hepatitis B	70.2	infectious diseases	6.18E-06
Neonatal bradycardia or tachycardia	656.9	pregnancy complications	6.3E-06
Rupture of synovium	727.5	musculoskeletal	6.54E-06
Osteoarthrosis	740	musculoskeletal	7.12E-06
Abnormal results of function study of kidney	589	genitourinary	7.7E-06
Tics and choreas	333.3	neurological	7.96E-06
Spinal stenosis of lumbar region	720.1	musculoskeletal	8.24E-06
Hematemesis	578.1	digestive	8.52E-06
Poisoning by agents primarily affecting blood constituents	964	injuries & poisonings	8.81E-06
Other nondiabetic retinopathy	362.3	sense organs	9.26E-06
Abnormal heart sounds	396	circulatory system	1.24E-05
Arthropathy associated with other disorders classified elsewhere	713	musculoskeletal	1.25E-05
Fracture of unspecified bones	809	injuries & poisonings	1.25E-05
Other and unspecified disorders of back	724	musculoskeletal	1.27E-05
Peripheral angiopathy in diseases classified elsewhere	443.7	circulatory system	1.3E-05
Other persistent mental disorders due to conditions classified elsewhere	290.3	mental disorders	1.36E-05
Gastritis and duodenitis, NOS	535.9	digestive	1.47E-05
Liver abscess and sequelae of chronic liver disease	571.8	digestive	1.47E-05
Mitral valve disease	394.2	circulatory system	1.5E-05
Viral infection	79	infectious diseases	1.52E-05
Other abnormality of urination	599.9	genitourinary	1.84E-05
Cerebral edema and compression of brain	348.2	neurological	1.9E-05
Ulceration of intestine	556.1	digestive	1.95E-05
Lung disease due to external agents	500	respiratory	2.2E-05
Diseases of spleen	289.5	hematopoietic	2.38E-05
Hypothyroidism NOS	244.4	endocrine/metabolic	2.41E-05
Ulceration of the lower GI tract	556	digestive	2.44E-05
Other disorders of soft tissues	729	musculoskeletal	2.47E-05
Orthopnea	513.32	respiratory	2.54E-05
Inflammatory diseases of female pelvic organs	614	genitourinary	2.64E-05
Colon cancer	153.2	neoplasms	2.85E-05

Description	Phecode	Group	p-value
Non-healing surgical wound	875	injuries & poisonings	2.99E-05
Esophageal bleeding (varices/hemorrhage)	530.2	digestive	2.99E-05

APPENDIX E

Single Nucleotide Polymorphism (SNP) – Phenotype Replications

SNP	PheCode	Description	OR	P-value
rs9271366	335	Multiple sclerosis	2.86	1.5E-23
rs3135388	335	Multiple sclerosis	2.86	9.6E-23
rs3129934	335	Multiple sclerosis	2.46	6.7E-18
rs6843082	427.21	Atrial fibrillation	1.41	4.8E-14
rs17042171	427.21	Atrial fibrillation	1.52	1.4E-13
rs2200733	427.21	Atrial fibrillation	1.52	1.9E-13
rs2200733	427.2	Atrial fibrillation and flutter	1.50	5.4E-13
rs7903146	250.2	Type 2 diabetes	1.23	2.6E-10
exm1615904	571.5	Other chronic nonalcoholic liver disease	1.59	3.7E-10
rs6910071	714.1	Rheumatoid arthritis	1.50	6.1E-10
rs3775948	274.1	Gout	0.65	6.2E-10
rs6855911	274.1	Gout	0.65	7.0E-10
rs6679677	250.1	Type 1 diabetes	2.10	8.2E-10
exm85427	250.1	Type 1 diabetes	2.09	9.9E-10
rs7442295	274.1	Gout	0.64	1.5E-09
rs734553	274.1	Gout	0.66	2.9E-09
rs12203592	172.2	Other non-epithelial cancer of skin	1.33	3.3E-09
exm389455	274.1	Gout	0.65	4.0E-09
rs2981579	174.11	Malignant neoplasm of female breast	1.33	5.1E-09
rs910873	172.11	Melanomas of skin	1.62	5.2E-09
rs3135338	335	Multiple sclerosis	1.76	6.5E-09
rs13129697	274.1	Gout	0.68	6.6E-09
rs6457617	714.1	Rheumatoid arthritis	1.41	1.0E-08
rs1219648	174.11	Malignant neoplasm of female breast	1.31	1.9E-08
rs2981575	174.11	Malignant neoplasm of female breast	1.31	2.1E-08
rs7901695	250.2	Type 2 diabetes	1.20	2.3E-08
rs157580	272.11	Hypercholesterolemia	0.83	2.5E-08
rs4506565	250.2	Type 2 diabetes	1.20	3.5E-08
rs11805303	555.1	Regional enteritis	1.84	8.2E-08
rs8034191	165.1	Cancer of bronchus; lung	1.32	1.4E-07
rs10993994	185	Cancer of prostate	1.33	1.9E-07
rs7130881	185	Cancer of prostate	1.42	2.6E-07
rs10737680	362.2	Degeneration of macula and posterior pole of retina	0.72	3.0E-07

SNP	PheCode	Description	OR	P-value
rs1410996	362.2	Degeneration of macula and posterior pole of retina	0.72	3.9E-07
rs2981582	174.11	Malignant neoplasm of female breast	1.28	4.1E-07
rs1329428	362.2	Degeneration of macula and posterior pole of retina	0.72	4.3E-07
rs16861990	452	Other venous embolism and thrombosis	1.54	6.2E-07
rs9268645	250.1	Type 1 diabetes	1.60	7.2E-07
rs965513	193	Thyroid cancer	1.43	1.4E-06
rs1016343	185	Cancer of prostate	1.35	2.1E-06
exm1272378	172.2	Other non-epithelial cancer of skin	1.36	2.8E-06
rs1333049	411	Ischemic Heart Disease	1.15	3.7E-06
rs8042374	165.1	Cancer of bronchus; lung	0.74	4.7E-06
rs10484554	696.4	Psoriasis	1.73	5.1E-06
rs1333049	411.4	Coronary atherosclerosis	1.16	5.8E-06
rs380390	362.2	Degeneration of macula and posterior pole of retina	1.32	6.9E-06
rs4977574	411	Ischemic Heart Disease	1.15	8.6E-06
rs13192471	714.1	Rheumatoid arthritis	1.41	1.1E-05
rs4785763	172.11	Melanomas of skin	1.28	1.5E-05
rs1329424	362.2	Degeneration of macula and posterior pole of retina	1.31	1.7E-05
rs4420638	272.11	Hypercholesterolemia	1.20	2.3E-05
rs1447295	185	Cancer of prostate	1.42	2.8E-05
rs687621	452	Other venous embolism and thrombosis	1.24	3.0E-05
rs505922	452	Other venous embolism and thrombosis	1.24	3.2E-05
rs3793917	362.2	Degeneration of macula and posterior pole of retina	1.33	3.6E-05
rs11228565	185	Cancer of prostate	1.30	3.9E-05
exm861570	362.2	Degeneration of macula and posterior pole of retina	1.33	4.2E-05
rs7837688	185	Cancer of prostate	1.41	4.2E-05
rs3123078	185	Cancer of prostate	1.25	4.5E-05
rs2131925	272.11	Hypercholesterolemia	0.87	5.8E-05
rs3802177	250.2	Type 2 diabetes	0.87	5.9E-05
exm717053	250.2	Type 2 diabetes	0.88	7.0E-05
rs258322	172.11	Melanomas of skin	1.40	7.5E-05
rs6983267	185	Cancer of prostate	0.80	7.8E-05
rs1512268	185	Cancer of prostate	1.25	7.9E-05
rs10889353	272.11	Hypercholesterolemia	0.87	8.0E-05
rs2000999	272.11	Hypercholesterolemia	1.18	8.1E-05

SNP	PheCode	Description	OR	P-value
rs4242382	185	Cancer of prostate	1.39	9.1E-05
rs4242384	185	Cancer of prostate	1.39	9.2E-05
rs445925	272.11	Hypercholesterolemia	0.80	9.5E-05
rs12740374	272.11	Hypercholesterolemia	0.85	9.9E-05
rs646776	272.11	Hypercholesterolemia	0.85	1.1E-04
rs629301	272.11	Hypercholesterolemia	0.85	1.2E-04
rs6679677	244	Hypothyroidism	1.22	1.2E-04
rs599839	272.11	Hypercholesterolemia	0.86	1.3E-04
rs9268853	714.1	Rheumatoid arthritis	1.26	1.6E-04
rs2660753	185	Cancer of prostate	1.36	2.0E-04
rs2412973	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.37	2.1E-04
rs7608910	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.38	2.3E-04
exm1327856	743	Osteoporosis, osteopenia and pathological fracture	0.88	2.3E-04
rs2954029	272.11	Hypercholesterolemia	0.88	2.5E-04
rs9349379	411.4	Coronary atherosclerosis	1.13	2.6E-04
rs1393350	172.11	Melanomas of skin	1.25	2.6E-04
rs4784227	174.11	Malignant neoplasm of female breast	1.22	2.7E-04
rs7111341	250.1	Type 1 diabetes	0.65	3.0E-04
rs10166942	340	Migraine	0.73	3.1E-04
rs2546890	696.4	Psoriasis	0.70	3.1E-04
rs401681	165.1	Cancer of bronchus; lung	0.83	3.2E-04
rs1558902	278.1	Obesity	1.15	3.5E-04
rs1421085	278.1	Obesity	1.15	3.6E-04
rs1150754	695.42	Systemic lupus erythematosus	1.54	4.4E-04
rs3131379	695.42	Systemic lupus erythematosus	1.60	4.5E-04
rs12654264	272.11	Hypercholesterolemia	1.13	4.9E-04
rs7578326	250.2	Type 2 diabetes	0.89	5.0E-04
rs7703051	272.11	Hypercholesterolemia	1.13	5.1E-04
rs3846662	272.11	Hypercholesterolemia	1.12	5.8E-04
rs4430796	185	Cancer of prostate	0.83	5.8E-04
rs1421085	278.11	Morbid obesity	1.23	5.8E-04
rs1558902	278.11	Morbid obesity	1.23	6.3E-04
rs5743289	555.1	Regional enteritis	1.56	6.3E-04
rs7517847	555	Inflammatory bowel disease and other gastroenteritis and colitis	0.74	6.5E-04
rs3846663	272.11	Hypercholesterolemia	1.13	6.5E-04

SNP	PheCode	Description	OR	P-value
rs1121980	278.11	Morbid obesity	1.23	6.7E-04
exm1037423	411	Ischemic Heart Disease	1.11	7.1E-04
rs3104767	327.71	Restless legs syndrome	0.67	7.1E-04
rs13073817	555.1	Regional enteritis	1.47	7.9E-04
rs2075650	272.11	Hypercholesterolemia	1.17	8.2E-04
rs9349379	411	Ischemic Heart Disease	1.11	8.3E-04
rs562338	272.11	Hypercholesterolemia	0.86	8.5E-04
rs3734805	174.11	Malignant neoplasm of female breast	1.32	8.7E-04
rs7517847	555.1	Regional enteritis	0.68	9.7E-04
rs17782313	278.1	Obesity	1.16	9.7E-04
rs9939609	278.1	Obesity	1.13	9.8E-04
exm1414617	362.2	Degeneration of macula and posterior pole of retina	1.27	9.9E-04
rs2237892	250.2	Type 2 diabetes	0.80	1.0E-03
rs8050136	278.1	Obesity	1.13	1.1E-03
rs266849	796	Elevated prostate specific antigen [PSA]	0.73	1.1E-03
rs17817449	278.1	Obesity	1.13	1.1E-03
rs964184	272.12	Hyperglyceridemia	1.62	1.2E-03
rs13376333	427.21	Atrial fibrillation	1.14	1.2E-03
rs964184	272.11	Hypercholesterolemia	1.17	1.2E-03
rs675209	274.1	Gout	1.21	1.3E-03
rs1265181	696.4	Psoriasis	1.41	1.3E-03
rs1121980	278.1	Obesity	1.13	1.3E-03
rs7756992	250.2	Type 2 diabetes	1.12	1.4E-03
rs31489	165.1	Cancer of bronchus; lung	0.84	1.4E-03
rs515135	272.11	Hypercholesterolemia	0.87	1.5E-03
rs9941349	278.11	Morbid obesity	1.21	1.5E-03
rs4689388	250.2	Type 2 diabetes	0.91	1.5E-03
rs10871777	278	Overweight, obesity and other hyperalimentation	1.14	1.6E-03
rs571312	278.1	Obesity	1.15	1.6E-03
rs3803662	174.11	Malignant neoplasm of female breast	1.18	1.7E-03
rs9465871	250.2	Type 2 diabetes	1.13	1.8E-03
rs4975616	165.1	Cancer of bronchus; lung	0.85	1.8E-03
rs1801214	250.2	Type 2 diabetes	0.91	1.9E-03
rs10440833	250.2	Type 2 diabetes	1.11	2.0E-03
rs2517532	244	Hypothyroidism	0.91	2.2E-03
rs4973768	174.11	Malignant neoplasm of female breast	1.16	2.3E-03
rs7574865	695.42	Systemic lupus erythematosus	1.40	2.4E-03

SNP	PheCode	Description	OR	P-value
rs6548238	278.1	Obesity	0.85	2.5E-03
rs7359397	278.1	Obesity	1.12	2.6E-03
rs10795668	153	Colorectal cancer	0.83	2.6E-03
rs1004446	250.1	Type 1 diabetes	0.74	2.6E-03
exm812431	281.1	Megaloblastic anemia	1.32	2.6E-03
rs4409764	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.29	2.8E-03
exm67254	555	Inflammatory bowel disease and other gastroenteritis and colitis	0.50	2.8E-03
rs1558902	278	Overweight, obesity and other hyperalimentation	1.11	2.8E-03
rs9939609	250.2	Type 2 diabetes	1.10	3.0E-03
rs2943641	250.2	Type 2 diabetes	0.91	3.1E-03
rs8050136	250.2	Type 2 diabetes	1.10	3.2E-03
rs1440072	278.1	Obesity	1.24	3.3E-03
rs11650066	411	Ischemic Heart Disease	1.10	3.6E-03
rs1840440	278.1	Obesity	0.89	3.6E-03
rs28927680	272.12	Hyperglyceridemia	1.72	4.2E-03
rs12286037	272.12	Hyperglyceridemia	1.71	4.2E-03
exm190281	574	Cholelithiasis and cholecystitis	1.35	4.2E-03
rs4402960	250.2	Type 2 diabetes	1.10	4.3E-03
rs1470579	250.2	Type 2 diabetes	1.10	4.4E-03
rs9870680	296.22	Major depressive disorder	1.21	4.5E-03
rs6465657	185	Cancer of prostate	1.17	4.5E-03
rs6769511	250.2	Type 2 diabetes	1.10	4.6E-03
exm1487912	281.1	Megaloblastic anemia	1.29	4.6E-03
rs7765379	714.1	Rheumatoid arthritis	1.28	4.9E-03
rs6504218	411	Ischemic Heart Disease	0.92	4.9E-03
rs2255141	272.11	Hypercholesterolemia	1.11	5.0E-03
rs402710	165.1	Cancer of bronchus; lung	0.86	5.0E-03
rs1999805	743	Osteoporosis, osteopenia and pathological fracture	1.09	5.3E-03
rs1701704	495	Asthma	1.16	5.3E-03
rs17582416	555.1	Regional enteritis	1.38	5.5E-03
exm1037423	714.1	Rheumatoid arthritis	1.18	5.6E-03
rs2237895	250.2	Type 2 diabetes	1.09	5.8E-03
rs9982601	411	Ischemic Heart Disease	1.13	5.9E-03
rs2187668	695.42	Systemic lupus erythematosus	1.43	6.1E-03
rs7593730	250.2	Type 2 diabetes	0.90	6.1E-03

SNP	PheCode	Description	OR	P-value
rs11708067	250.2	Type 2 diabetes	0.91	6.5E-03
rs944289	193	Thyroid cancer	0.82	6.7E-03
rs2076756	555.1	Regional enteritis	1.38	6.7E-03
rs492602	281.1	Megaloblastic anemia	1.28	6.8E-03
rs2867125	278.1	Obesity	0.87	6.8E-03
rs10896449	185	Cancer of prostate	0.86	6.9E-03
rs5743289	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.32	6.9E-03
rs9642880	189.21	Malignant neoplasm of bladder	1.26	7.0E-03
rs10965250	250.2	Type 2 diabetes	0.89	7.1E-03
rs2112347	278.11	Morbid obesity	0.84	7.1E-03
rs7865618	411	Ischemic Heart Disease	0.92	7.2E-03
rs17085007	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.33	7.3E-03
exm893239	250.2	Type 2 diabetes	1.09	7.3E-03
rs2306374	411	Ischemic Heart Disease	1.12	7.4E-03
rs704010	174.11	Malignant neoplasm of female breast	1.14	7.4E-03
rs909116	174.11	Malignant neoplasm of female breast	0.88	7.6E-03
rs4938303	272.12	Hyperglyceridemia	1.41	7.6E-03
rs5754217	695.42	Systemic lupus erythematosus	1.36	7.8E-03
rs12970134	278.1	Obesity	1.12	7.8E-03
rs1165205	274.1	Gout	0.87	7.9E-03
rs2076756	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.28	8.0E-03
exm2263569	272.11	Hypercholesterolemia	0.87	8.4E-03
rs1701704	250.1	Type 1 diabetes	1.29	8.4E-03
rs614367	174.11	Malignant neoplasm of female breast	1.19	8.5E-03
rs9284813	185	Cancer of prostate	1.22	8.8E-03
rs987870	495	Asthma	1.19	9.0E-03
rs11206510	411	Ischemic Heart Disease	0.90	9.5E-03
rs651007	272.11	Hypercholesterolemia	1.11	9.5E-03
rs987237	278.1	Obesity	1.13	9.8E-03
rs801114	172.21	Basal cell carcinoma	1.30	9.9E-03
rs401681	172.21	Basal cell carcinoma	0.77	9.9E-03
rs2292239	250.1	Type 1 diabetes	1.28	1.0E-02
rs6831256	272.11	Hypercholesterolemia	1.09	1.0E-02
rs9930506	278.1	Obesity	1.10	1.0E-02
rs2019960	335	Multiple sclerosis	1.32	1.0E-02
rs10883365	555.1	Regional enteritis	1.34	1.0E-02

SNP	PheCode	Description	OR	P-value
rs4409764	555.1	Regional enteritis	1.34	1.0E-02
rs12678919	272.12	Hyperglyceridemia	0.52	1.0E-02
rs10503669	272.12	Hyperglyceridemia	0.52	1.0E-02
rs7754840	250.2	Type 2 diabetes	1.09	1.1E-02
exm282443	555	Inflammatory bowel disease and other gastroenteritis and colitis	1.30	1.1E-02
rs10927875	425.1	Primary/intrinsic cardiomyopathies	0.85	1.1E-02
rs10946398	250.2	Type 2 diabetes	1.09	1.1E-02
rs1530440	401	Hypertension	0.92	1.2E-02
rs11668477	272.11	Hypercholesterolemia	0.90	1.2E-02
rs1332844	411	Ischemic Heart Disease	0.92	1.2E-02
rs10486567	185	Cancer of prostate	0.85	1.2E-02
rs4712524	250.2	Type 2 diabetes	1.09	1.2E-02
rs17482753	272.12	Hyperglyceridemia	0.54	1.2E-02
rs2523393	335	Multiple sclerosis	0.78	1.2E-02
exm686341	272.12	Hyperglyceridemia	0.54	1.2E-02
rs10401969	272.11	Hypercholesterolemia	0.85	1.2E-02
rs16996148	272.11	Hypercholesterolemia	0.86	1.2E-02
rs2282679	261.4	Vitamin D deficiency	1.14	1.2E-02
exm1495645	796	Elevated prostate specific antigen [PSA]	0.69	1.2E-02
rs228769	743	Osteoporosis, osteopenia and pathological fracture	0.91	1.2E-02
rs7023329	172.11	Melanomas of skin	1.15	1.2E-02
rs2206277	278.1	Obesity	1.13	1.2E-02
rs6029526	272.11	Hypercholesterolemia	1.09	1.2E-02
rs180730	250.4	Abnormal glucose	1.16	1.2E-02
rs1799884	250.4	Abnormal glucose	1.17	1.3E-02
rs1008953	696.4	Psoriasis	0.73	1.3E-02
rs635634	272.11	Hypercholesterolemia	1.11	1.3E-02
rs4712523	250.2	Type 2 diabetes	1.09	1.3E-02
rs2274089	280.1	Iron deficiency anemias, unspecified or not due to blood loss	0.78	1.3E-02
rs7957197	250.2	Type 2 diabetes	0.91	1.3E-02
rs1335532	335	Multiple sclerosis	0.66	1.3E-02
rs12280753	272.12	Hyperglyceridemia	1.59	1.3E-02
rs11024074	401	Hypertension	1.08	1.3E-02
rs445	288.2	Elevated white blood cell count	0.78	1.3E-02
rs12531711	695.42	Systemic lupus erythematosus	1.42	1.3E-02
rs10488631	695.42	Systemic lupus erythematosus	1.42	1.3E-02

SNP	PheCode	Description	OR	P-value
rs8102137	189.21	Malignant neoplasm of bladder	1.25	1.3E-02
rs1183201	274.1	Gout	0.87	1.3E-02
rs12446956	296.22	Major depressive disorder	0.76	1.4E-02
rs4607517	250.4	Abnormal glucose	1.17	1.4E-02
rs11190140	555.1	Regional enteritis	1.32	1.4E-02
rs935334	401	Hypertension	1.10	1.4E-02
rs10788160	796	Elevated prostate specific antigen [PSA]	1.21	1.4E-02
rs2121070	401	Hypertension	1.09	1.4E-02
rs2207418	416	Cardiomegaly	1.14	1.4E-02
rs610604	696.4	Psoriasis	1.28	1.5E-02
rs10484561	202.21	Nodular lymphoma	1.49	1.5E-02
rs9271366	555.1	Regional enteritis	0.64	1.5E-02
exm1272378	172.21	Basal cell carcinoma	1.47	1.6E-02
exm236837	250.1	Type 1 diabetes	0.79	1.6E-02
exm572471	696.4	Psoriasis	1.47	1.6E-02
rs2618476	695.42	Systemic lupus erythematosus	1.29	1.7E-02
rs10757278	411.2	Myocardial infarction	1.13	1.7E-02
exm521729	280.1	Iron deficiency anemias, unspecified or not due to blood loss	0.76	1.7E-02
rs8170	174.11	Malignant neoplasm of female breast	1.16	1.8E-02
rs6511720	272.11	Hypercholesterolemia	0.88	1.8E-02
rs925946	278.1	Obesity	1.10	1.8E-02
rs4977574	411.2	Myocardial infarction	1.12	1.8E-02
rs7543130	442.1	Aortic aneurysm	0.84	2.0E-02
rs718314	189.11	Malignant neoplasm of kidney, except pelvis	1.24	2.0E-02
rs12748152	272.11	Hypercholesterolemia	1.15	2.1E-02
rs2941740	743	Osteoporosis, osteopenia and pathological fracture	0.93	2.1E-02
rs7927997	555.1	Regional enteritis	1.30	2.1E-02
rs1531343	250.2	Type 2 diabetes	1.12	2.1E-02
rs6859219	714.1	Rheumatoid arthritis	0.84	2.3E-02
rs9533090	743	Osteoporosis, osteopenia and pathological fracture	1.07	2.3E-02
rs541862	362.2	Degeneration of macula and posterior pole of retina	0.77	2.4E-02
rs2300747	335	Multiple sclerosis	0.69	2.4E-02
rs1564348	272.11	Hypercholesterolemia	1.11	2.4E-02
rs9478751	274.1	Gout	1.15	2.8E-02
rs3810291	278.1	Obesity	0.92	2.8E-02

SNP	**PheCode**	**Description**	**OR**	**P-value**
rs16948048	401	Hypertension	1.06	2.9E-02
rs9652490	333.1	Essential tremor	1.32	3.0E-02
rs7501939	185	Cancer of prostate	0.88	3.1E-02
rs6687758	153	Colorectal cancer	1.16	3.2E-02
rs13277113	695.42	Systemic lupus erythematosus	1.26	3.4E-02
rs13003464	555.1	Regional enteritis	1.27	3.6E-02
rs2081687	272.11	Hypercholesterolemia	1.08	4.0E-02
rs6882076	272.11	Hypercholesterolemia	0.93	4.1E-02
rs1501908	272.11	Hypercholesterolemia	0.93	4.1E-02
rs1562430	174.11	Malignant neoplasm of female breast	0.91	4.4E-02
rs10852932	442.1	Aortic aneurysm	1.17	4.5E-02
rs4245791	272.11	Hypercholesterolemia	1.07	4.9E-02
rs4299376	272.11	Hypercholesterolemia	1.07	4.9E-02

APPENDIX F

Significant SNP-Phenotype Interactions

SNP	PheCode	Description	SNP OR	SNP P-value	Smoking P-value	Interaction P-value
rs10484561	202.21	Nodular lymphoma	1.49	2.4E-05	4.5E-03	4.1E-05
rs2621416	202.21	Nodular lymphoma	1.20	1.9E-03	8.7E-03	9.8E-04
rs1000113	555.1	Regional enteritis	1.43	3.7E-02	1.9E-01	1.7E-02
rs3024505	555.1	Regional enteritis	1.05	7.1E-03	1.1E-01	8.2E-03
rs11747270	555.1	Regional enteritis	1.33	6.7E-02	1.9E-01	1.9E-02
rs7714584	555.1	Regional enteritis	1.33	6.7E-02	1.9E-01	1.9E-02
rs13361189	555.1	Regional enteritis	1.37	6.5E-02	2.0E-01	2.2E-02
rs4846914	272.12	Hyperglyceridemia	1.26	1.4E-04	2.9E-02	1.5E-03
rs2144300	272.12	Hyperglyceridemia	1.26	1.5E-04	3.0E-02	1.6E-03
rs11101442	695.42	Systemic lupus erythematosus	1.00	5.1E-02	5.2E-02	8.7E-03
exm823419	695.42	Systemic lupus erythematosus	0.89	8.9E-01	9.2E-02	2.2E-02
exm572471	696.4	Psoriasis	1.47	1.1E-01	2.6E-02	3.4E-02
rs425105	250.1	Type 1 diabetes	0.81	3.3E-01	2.6E-02	2.1E-02
rs17388568	250.1	Type 1 diabetes	1.01	2.9E-02	2.3E-01	1.2E-02
rs445	288.2	Elevated white blood cell count	0.78	4.7E-01	9.9E-01	8.5E-03
rs3117582	165.1	Cancer of bronchus; lung	0.93	4.3E-01	5.6E-01	4.2E-03
rs987870	495	Asthma	1.19	1.9E-03	8.6E-01	9.6E-03
rs6859219	714.1	Rheumatoid arthritis	0.84	3.0E-01	8.6E-01	2.8E-02
rs614367	174.11	Malignant neoplasm of female breast	1.19	2.2E-02	5.3E-01	4.4E-02
rs541862	362.2	Degeneration of macula and posterior pole of retina	0.77	9.9E-01	2.0E-01	4.4E-02
rs2681472	401	Hypertension	0.97	1.3E-01	2.9E-01	4.1E-03
rs2681492	401	Hypertension	0.97	1.6E-01	8.9E-01	4.2E-03
rs17249754	401	Hypertension	0.97	2.1E-01	8.9E-01	4.7E-03
rs17609240	288.2	Elevated white blood cell count	0.93	3.0E-01	5.9E-13	7.2E-03
rs8139900	274.1	Gout	0.97	3.1E-02	2.4E-02	1.3E-02
rs5759167	185	Cancer of prostate	1.09	1.8E-02	5.1E-02	1.7E-02
rs4876662	442.1	Aortic aneurysm	1.17	2.5E-04	3.2E-140	2.3E-02

SNP	PheCode	Description	SNP OR	SNP P-value	Smoking P-value	Interaction P-value
rs17030613	401	Hypertension	1.00	1.1E-01	1.0E-09	6.5E-03
rs7237848	288.2	Elevated white blood cell count	1.10	1.5E-01	2.3E-01	2.1E-02
rs6504218	411	Ischemic Heart Disease	0.92	5.1E-01	4.9E-02	1.5E-03
rs2943641	250.2	Type 2 diabetes	0.91	3.9E-01	8.5E-02	1.0E-02
rs2081687	272.11	Hypercholesterolemia	1.08	4.0E-03	7.0E-01	1.2E-02
rs10458787	278.1	Obesity	0.94	8.8E-01	6.2E-02	3.8E-02
rs579459	411	Ischemic Heart Disease	1.06	1.4E-01	2.1E-01	3.8E-02
rs10933436	411	Ischemic Heart Disease	1.02	1.5E-01	5.8E-02	1.0E-02
rs10906115	250.2	Type 2 diabetes	1.01	4.9E-01	3.4E-01	1.4E-02
rs11646213	401	Hypertension	0.97	2.1E-01	3.6E-01	3.0E-02
rs7865618	411	Ischemic Heart Disease	0.92	8.5E-03	3.6E-01	1.9E-02
rs12946454	401	Hypertension	1.06	4.9E-01	9.4E-01	2.2E-02
rs1038304	743	Osteoporosis, osteopenia and pathological fracture	0.97	2.9E-02	9.1E-01	1.4E-02
rs12518099	250.2	Type 2 diabetes	0.94	2.1E-01	9.2E-01	1.3E-02
rs402710	165.1	Cancer of bronchus; lung	0.86	4.0E-03	7.5E-05	1.3E-02
rs4743150	411	Ischemic Heart Disease	1.00	3.3E-02	5.3E-02	3.0E-03
rs4026608	442.1	Aortic aneurysm	1.13	8.6E-01	1.5E-06	3.4E-02
rs2517532	244	Hypothyroidism	0.91	7.8E-04	5.1E-11	4.2E-03
rs17608766	401	Hypertension	0.98	3.3E-03	2.5E-01	4.5E-03
rs4785763	172.11	Melanomas of skin	1.28	1.3E-01	3.8E-01	4.5E-02
rs3095254	288.2	Elevated white blood cell count	1.04	9.4E-01	2.9E-01	7.9E-03
rs7501939	185	Cancer of prostate	0.88	1.2E-01	4.1E-01	1.9E-02
rs2383207	442.11	Abdominal aortic aneurysm	0.98	2.5E-01	9.4E-01	8.4E-03
rs16951095	165.1	Cancer of bronchus; lung	1.07	4.5E-01	6.5E-04	9.1E-03
rs7626795	165.1	Cancer of bronchus; lung	0.91	2.9E-02	4.2E-01	1.3E-03
rs10497721	250.2	Type 2 diabetes	1.01	9.8E-02	2.6E-01	2.3E-03

SNP	**PheCode**	**Description**	**SNP OR**	**SNP P-value**	**Smoking P-value**	**Interactio n P-value**
rs4415084	174.11	Malignant neoplasm of female breast	1.01	1.1E-01	6.8E-02	1.3E-02
rs229541	250.1	Type 1 diabetes	1.13	9.2E-01	1.7E-03	3.1E-02
rs3741208	250.1	Type 1 diabetes	1.11	6.3E-01	7.2E-02	2.2E-02
rs1926657	174.11	Malignant neoplasm of female breast	1.02	2.1E-01	7.0E-02	1.4E-02

REFERENCES

1. Jemal, A. *et al.* Global cancer statistics. *CA. Cancer J. Clin.* **61,** 69–90 (2011).

2. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2012. *CA. Cancer J. Clin.* **62,** 10–29 (2012).

3. Cancer Statistics Review, 1975-2011 - Previous Version - SEER Cancer Statistics Review. Available at: https://seer.cancer.gov/archive/csr/1975_2011/. (Accessed: 30th May 2017)

4. Goldstraw, P. *et al.* The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **2,** 706–714 (2007).

5. Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer. *N. Engl. J. Med.* **368,** 1980–1991 (2013).

6. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N. Engl. J. Med.* **365,** 395–409 (2011).

7. Moyer, V. A. Screening for Lung Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **160,** 330–338 (2014).

8. Lewis, J. A. *et al.* Low-Dose CT Lung Cancer Screening Practices and Attitudes among Primary Care Providers at an Academic Medical Center. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **24,** 664–670 (2015).

9. Atlas, S. J. *et al.* A Cluster-Randomized Trial of a Primary Care Informatics-Based System for Breast Cancer Screening. *J. Gen. Intern. Med.* **26,** 154–161 (2011).

10. Chaudhry R, Scheitel SM, McMurtry EK & et al. Web-based proactive system to improve breast cancer screening: A randomized controlled trial. *Arch. Intern. Med.* **167,** 606–611 (2007).

11. White, P. & Kenton, K. Use of Electronic Medical Record–Based Tools to Improve Compliance With Cervical Cancer Screening Guidelines: Effect of an Educational Intervention on Physicians' Practice Patterns. *J. Low. Genit. Tract Dis.* **17,** 175–181 (2013).

12. Dupuis, E. A. *et al.* Tracking Abnormal Cervical Cancer Screening: Evaluation of an EMR-Based Intervention. *J. Gen. Intern. Med.* **25,** 575–580 (2010).

13. White, P. Effects of Electronic Health Record–Based Interventions on Cervical Cancer Screening in Adolescents: A 1-Year Follow-up. *J. Low. Genit. Tract Dis.* **18,** 169–173 (2014).

14. Schroy, P. C. *et al.* The Impact of a Novel Computer-Based Decision Aid on Shared Decision-Making for Colorectal Cancer Screening: A Randomized Trial (Running head: SDM for CRC Screening). *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.* **31,** 93–107 (2011).

15. Ruffin IV, M. T., Fetters, M. D. & Jimbo, M. Preference-based electronic decision aid to promote colorectal cancer screening: Results of a randomized controlled trial. *Prev. Med.* **45,** 267–273 (2007).

16. Humphrey, L. L. *et al.* Improving the Follow-Up of Positive Hemoccult Screening Tests: An Electronic Intervention. *J. Gen. Intern. Med.* **26,** 691–697 (2011).

17. Nease, D. E. *et al.* Impact of a Generalizable Reminder System on Colorectal Cancer Screening in Diverse Primary Care Practices. *Med. Care* **46,** S68–S73 (2008).

18. Sequist, T. D., Zaslavsky, A. M., Colditz, G. A. & Ayanian, J. Z. Electronic Patient Messages to Promote Colorectal Cancer Screening: A Randomized, Controlled Trial. *Arch. Intern. Med.* **171,** 636–641 (2011).

19. Miller, D. P. *et al.* Effectiveness of a Web-Based Colorectal Cancer Screening Patient Decision Aid. *Am. J. Prev. Med.* **40,** 608–615 (2011).

20. Ornstein, S., Nemeth, L. S., Jenkins, R. G. & Nietert, P. J. Colorectal Cancer Screening in Primary Care: Translating Research into Practice. *Med. Care* **48,** 900–906 (2010).

21. Smith, S. K. *et al.* A decision aid to support informed choices about bowel cancer screening among adults with low education: randomised controlled trial. *BMJ* **341,** (2010).

22. Paluch-Shimon, S. *et al.* Prevention and screening in BRCA mutation carriers and other breast/ovarian hereditary cancer syndromes: ESMO Clinical Practice Guidelines for cancer prevention and screening. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **27,** v103–v110 (2016).

23. Wang, Y., Chen, E. S., Pakhomov, S., Lindemann, E. & Melton, G. B. Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record. *AMIA. Annu. Symp. Proc.* **2016,** 1209–1218 (2017).

24. Manning, C., Raghavan, P. & Schuetze, H. *Introduction to Information Retrieval*. (Cambridge University Press, 2008).

25. Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **2,** 113–124 (1956).

26. Chomsky, N. On certain formal properties of grammars. *Inf. Control* **2,** 137–167 (1959).

27. Kernighan, B. W. & Pike, R. *The Unix Programming Environment*. (Prentice-Hall, 1983).

28. Sager, N. Syntactic Analysis of Natural Language. *Adv. Comput.* **8,** 153–188 (1967).

29. Grishman, R., Sager, N., Raze, C. & Bookchin, B. The Linguistic String Parser. in *Proceedings of the June 4-8, 1973, National Computer Conference and Exposition* 427–434 (ACM, 1973). doi:10.1145/1499586.1499693

30. Sager, N. Sublanguage grammers in science information processing. *J. Am. Soc. Inf. Sci.* **26,** 10–16 (1975).

31. Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **2,** 121–167 (1998).

32. Eddy, S. R. What is a hidden Markov model? *Nat. Biotechnol.* **22,** 1315–1316 (2004).

33. Gene prediction with conditional random fields (PDF Download Available). *ResearchGate* Available at: https://www.researchgate.net/publication/228639471_Gene_prediction_with_conditional_random_fields. (Accessed: 10th June 2017)

34. Manning, C. D. & Schütze, H. *Foundations of Statistical Natural Language Processing*. (The MIT Press, 1999).

35. Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V. GATE: An Architecture for Development of Robust HLT Applications. in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* 168–175 (Association for Computational Linguistics, 2002). doi:10.3115/1073083.1073112

36. Ferrucci, D. & Lally, A. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Nat Lang Eng* **10,** 327–348 (2004).

37. Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. & Johnson, S. B. A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc. JAMIA* **1,** 161–174 (1994).

38. Aronson, A. R. & Lang, F.-M. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 229–236 (2010).

39. Savova, G., Kipper-Schuler, K., Buntrock, J. & Chute, C. Towards enhanced interoperability for large HLT systems: UIMA for NLP2008.

40. Denny, J. C. *et al.* Identifying UMLS concepts from ECG Impressions using KnowledgeMap. *AMIA. Annu. Symp. Proc.* **2005,** 196–200 (2005).

41. Rinaldi, F. *et al.* OntoGene in BioCreative II. *Genome Biol.* **9,** S13 (2008).

42. Rinaldi, F., Schneider, G. & Clematide, S. Relation mining experiments in the pharmacogenomics domain. *J. Biomed. Inform.* **45,** 851–861 (2012).

43. Rinaldi, F. *et al.* OntoGene web services for biomedical text mining. *BMC Bioinformatics* **15,** S6 (2014).

44. Liu, M. *et al.* A Study of Transportability of an Existing Smoking Status Detection Module across Institutions. *AMIA. Annu. Symp. Proc.* **2012,** 577–586 (2012).

45. Uzuner, O., Szolovits, P. & Kohane, I. i2b2 Workshop on Natural Language Processing Challenges for Clinical Records.

46. Uzuner, Ö., Goldstein, I., Luo, Y. & Kohane, I. Identifying Patient Smoking Status from Medical Discharge Records. *J. Am. Med. Inform. Assoc. JAMIA* **15,** 14–24 (2008).

47. Clark, C. *et al.* Identifying Smokers with a Medical Extraction System. *J. Am. Med. Inform. Assoc. JAMIA* **15,** 36–39 (2008).

48. Sohn, S. & Savova, G. K. Mayo clinic smoking status classification system: extensions and improvements. *AMIA Annu. Symp. Proc. AMIA Symp.* **2009,** 619–623 (2009).

49. Open Health Natural Language Processing (OHNLP) Consortium. Available at: http://www.ohnlp.org/index.php/Main_Page. (Accessed: 10th June 2017)

50. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84,** 362–369 (2008).

51. Zeng, Q. T. *et al.* Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* **6,** 30 (2006).

52. Himes, B. E., Kohane, I. S., Ramoni, M. F. & Weiss, S. T. Characterization of Patients who Suffer Asthma Exacerbations using Data Extracted from Electronic Medical Records. *AMIA. Annu. Symp. Proc.* **2008,** 308–312 (2008).

53. Xiaohua Zhou, Hyoil Han, Chankai, I., Prestrud, A. A. & Brooks, A. D. Converting Semi-structured Clinical Medical Records into Information and Knowledge. in 1162–1162 (IEEE, 2005). doi:10.1109/ICDE.2005.207

54. De Silva, L. *et al.* Extraction and Quantification of Pack-years and Classification of Smoker Information in Semi-structured Medical Records. *Proc. 28 Th Int. Conf. Mach. Learn. Bellevue WA USA* (2011).

55. Medicine, I. of, Practice, B. on P. H. and P. H. & Records, C. on the R. S. and B. D. and M. for E. H. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. (National Academies Press, 2015).

56. Wang, L., Ruan, X., Yang, P. & Liu, H. Comparison of Three Information Sources for Smoking Information in Electronic Health Records. *Cancer Inform.* **15,** 237–242 (2016).

57. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367 (2009).

58. Ha, N.-T., Freytag, S. & Bickeboeller, H. Coverage and efficiency in current SNP chips. *Eur. J. Hum. Genet.* **22,** 1124–1130 (2014).

59. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

60. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26,** 1205–1210 (2010).

61. McCarty, C. A. *et al.* The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4,** 13 (2011).

62. Warner, J. L. & Alterovitz, G. Phenome Based Analysis as a Means for Discovering Context Dependent Clinical Reference Ranges. *AMIA. Annu. Symp. Proc.* **2012,** 1441–1449 (2012).

63. Verma, A. *et al.* INTEGRATING CLINICAL LABORATORY MEASURES AND ICD-9 CODE DIAGNOSES IN PHENOME-WIDE ASSOCIATION STUDIES. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **21,** 168–179 (2016).

64. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30,** 2375–2376 (2014).

65. Bochenek, J. *exact_stats: A python module for doing bayesian statistics for case-control studies*. (PheWAS, 2014).

66. Flint, J. & Mackay, T. F. C. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* **19,** 723–733 (2009).

67. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).

68. Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343,** 1246980 (2014).

69. Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1204–1209 (2012).

70. Holmes, M. V. *et al.* A systematic review and meta-analysis of 130,000 individuals shows smoking does not modify the association of APOE genotype on risk of coronary heart disease. *Atherosclerosis* **237,** 5–12 (2014).

71. Polfus, L. M. *et al.* Genome-wide association study of gene by smoking interactions in coronary artery calcification. *PloS One* **8,** e74642 (2013).

72. Skare, Ø. *et al.* Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. *Ann. Hum. Genet.* **76,** 221–236 (2012).

73. Tyrrell, J. *et al.* Gene–obesogenic environment interactions in the UK Biobank study. *Int. J. Epidemiol.* **46,** 559–575 (2017).

74. Haldane, J. B. S. The Interaction of Nature and Nurture. *Ann. Eugen.* **13,** 197–205 (1946).

75. McGue, M. & Carey, B. E. Gene–Environment Interaction in the Behavioral Sciences: Findings, Challenges, and Prospects. in *Gene-Environment Transactions in Developmental Psychopathology* (eds. Tolan, P. H. & Leventhal, B. L.) 35–57 (Springer International Publishing, 2017). doi:10.1007/978-3-319-49227-8_3

76. Cuevas, J. *et al.* Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. *Plant Genome* **9,** (2016).

77. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **advance online publication,** (2017).

78. Clark, C. *et al.* Identifying Smokers with a Medical Extraction System. *J. Am. Med. Inform. Assoc. JAMIA* **15,** 36–39 (2008).

79. Clark, C. *et al.* MITRE system for clinical assertion status classification. *J. Am. Med. Inform. Assoc. JAMIA* **18,** 563–567 (2011).

80. Denny, J. C. *et al.* Evaluation of a method to identify and categorize section headers in clinical documents. *J. Am. Med. Inform. Assoc. JAMIA* **16,** 806–815 (2009).

81. Osterman, T. J. Quantifying Tobacco Exposure Using Clinical Notes and Natural Language Processing to Enable Lung Cancer Screening. (2015).

82. Apache License, Version 2.0. Available at: https://www.apache.org/licenses/LICENSE-2.0. (Accessed: 19th June 2017)

83. Jemal, A. & Fedewa, S. A. Lung Cancer Screening With Low-Dose Computed Tomography in the United States—2010 to 2015. *JAMA Oncol.* (2017). doi:10.1001/jamaoncol.2016.6416

84. Coffin, J., Duffie, C. & Furno, M. The Patient-Centered Medical Home and Meaningful Use: a challenge for better care. *J. Med. Pract. Manag. MPM* **29,** 331–334 (2014).

85. Mondul, A. M. *et al.* Association of serum α-tocopherol with sex steroid hormones and interactions with smoking: implications for prostate cancer risk. *Cancer Causes Control CCC* **22,** 827–836 (2011).

86. GWAS Catalog. Available at: https://www.ebi.ac.uk/gwas/home. (Accessed: 12th July 2017)

87. Samani, N. J. *et al.* Genomewide Association Analysis of Coronary Artery Disease. *N. Engl. J. Med.* **357,** 443–453 (2007).

88. Li, X. *et al.* Meta-analysis identifies robust association between SNP rs17465637 in MIA3 on chromosome 1q41 and coronary artery disease. *Atherosclerosis* **231,** 136–140 (2013).

89. Wang, A. Z., Li, L., Zhang, B., Shen, G.-Q. & Wang, Q. K. Association of SNP rs17465637 on chromosome 1q41 and rs599839 on 1p13.3 with Myocardial Infarction in an American Caucasian Population. *Ann. Hum. Genet.* **75,** 475–482 (2011).

90. Young, K. L. *et al.* Interaction of smoking and obesity susceptibility loci on adolescent BMI: The National Longitudinal Study of Adolescent to Adult Health. *BMC Genet.* **16,** (2015).

91. Zheng, J.-S. *et al.* Modulation by Dietary Fat and Carbohydrate of IRS1 Association With Type 2 Diabetes Traits in Two Populations of Different Ancestries. *Diabetes Care* **36,** 2621–2627 (2013).